Conversational Search with Large Language Models

Cumulative Habilitation Thesis
to obtain the license to teach in the subject Business Informatics
(Wirtschaftsinformatik)

Svitlana Vakulenko

Department of Information Systems and Operations Management
Vienna University of Economics and Business

Dated: January 7, 2026

Contents

I Preface

II Publications

Part I

Preface

1. Introduction

Information Retrieval (IR) is a scientific discipline that studies the representation, organization, storage, and retrieval of information [42, 4, 19, 9]. IR research is of utmost importance for any one of us, as it focuses on improving methods for scaling access to the totality of all human knowledge. Humanity has long surpassed the point at which the accumulated body of knowledge can be realistically processed by any single individual within their lifetime. At this stage, the primary bottleneck to effective information access is no longer storage, but retrieval.

Research in IR covers the design, implementation, evaluation, and analysis of systems that find and rank relevant information so as to satisfy a user's information need, often expressed as a search query. This typically implies practical applications to large collections of mostly unstructured data such as text documents. As shown in Figure 1.1, IR systems rank documents according to estimated relevance in response to a query. The most common example of a classic IR system is a web search engine, such as Google or Bing, that billions of people use every day.

The term "information retrieval" was first coined by Calvin N. Mooers in a 1950 technical report published by his company Zator Co. [22]. His definition of IR pointed out its main purpose of "finding information whose location or very existence is a priori unknown," and motivated the need for encoding and indexing documents to be able to locate this information.

The conceptual foundation for an information retrieval (IR) system was proposed even earlier, in 1945, by Vannevar Bush, an American engineer [8]. Bush imagined a theoretical machine, which he called Memex, that used a microfilm to store information and retrieved it via associative trails that provided links between related documents. These trails were to appear naturally as readers navigated through the documents. Memex has served as an inspiration for IR researchers ever since.

Over the past years, IR delivered scalable search systems that provide immediate access to billions of web pages for people all over the world within just milliseconds. The most striking, however, is that the modern web search systems still heavily rely on the models introduced in 1970s: inverted index and BM25. An inverted index is the core data structure of IR systems, enabling efficient search over large document collections [17, 18, 31, 32]. It maps terms to the documents in which they occur. By measuring the overlap between query terms and document terms, an IR system can estimate the relevance of a document to a given query. Retrieval models, such as TF-IDF and BM25, operationalize this idea by weighting terms according to their importance in individual documents (Term Frequency) and in the document collection as a whole (Inverse Document Frequency), which enables more accurate document ranking [27, 28].

In parallel to the development of main-stream search systems that assumed independent keyword queries as input, an alternative paradigm emerged in 1980s that advocated for modeling IR as an iterative process (interactive IR) [6, 5]. These ideas were later connected to research on Natural Language Processing (NLP) and dialogue systems [44], and lead the IR community to establish a new term of "conversational search" in 2017 [25, 7] that advocated for the future search systems with dialogue-based interfaces.

Conversational search aims to support design of an IR system that supports natural, multi-turn dialogue able to correctly maintain context over the course of an interaction to refine and fulfill information needs [25, 2]. Current research connects foundational IR concepts with emerging language modeling techniques to enable conversational search [21].

The wide adoption of Large Language Models (LLMs), that skyrocketed around the end of 2022

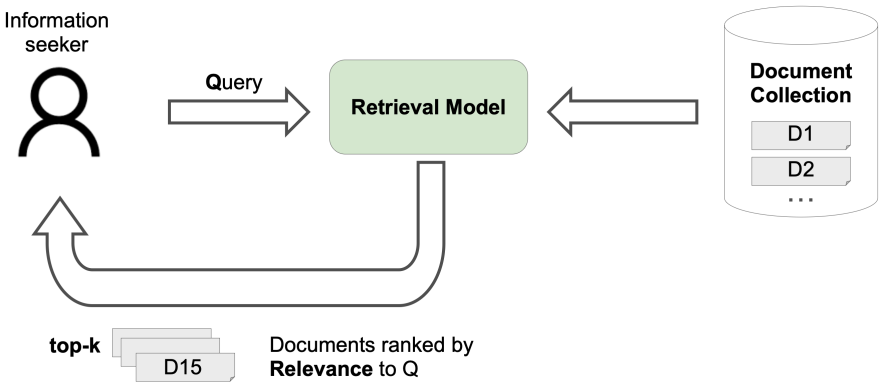


Figure 1.1: A high-level overview of a classic information retrieval (IR) system.

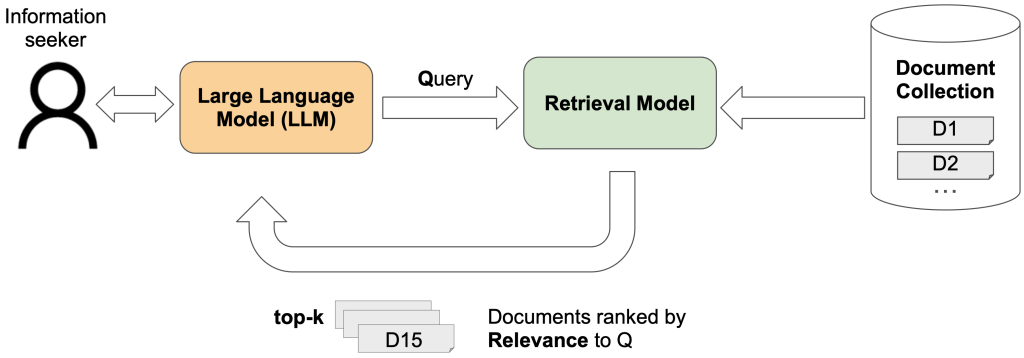


Figure 1.2: An overview of Retrieval Augmented Generation (RAG) framework.

due to the release of ChatGPT by OpenAI, has changed the way many people access information today and likely to have a lasting effect in the future. LLMs are machine-learning models trained on massive amounts of text data [43, 20]. They are designed to generate natural language by predicting sequences of text.

LLMs provide a natural-language interface to web knowledge and allow information seekers to better express their information needs in the course of a dialogue interaction. Since LLMs do not store the original web pages that they are trained on, they have to rely on (web) search engines to trace back to the original documents and access up-to-date information in real time. The current state-of-the-art LLM-based systems rely on Retrieval Augmented Generation (RAG) framework that combines an LLM with an IR model to improve accuracy and factuality of generated text [15]. With RAG, a search engine can serve as a tool for an LLM to request additional information in the middle of a dialogue interaction allowing the LLM to retrieve and incorporate relevant information in real time. As shown in Figure 1.2, an LLM can formulate a search query based on the dialogue context and then use the top-ranked documents returned by the IR model to generate a grounded and contextually-relevant response.

To successfully incorporate relevant information provided by RAG into dialogue context, an LLM has to correctly interpret the dialogue context as well as the retrieved documents. In practice, language models often struggle with both of these challenges. Two of the papers in this dissertation contain studies that introduce benchmark datasets demonstrating these limitations in practice and allowing to track progress on two complex tasks that require accurate natural language understanding: disambiguation and summarization.

One of the core tasks in dialogue is question answering (QA) [26] that requires correct question interpretation based on the dialogue context. My most prominent contribution to date, which is included in this habilitation thesis, is the subtask of question rewriting as an intermediate step in conversational QA that allows to separately assess question interpretation from the retrieval performance. This work

was a direct precursor to the current RAG approaches that integrate external search engines as tools via APIs to retrieve answers using rewritten conversational questions, as shown in Figure 1.2.

Last but not least, an information-seeking dialogue requires mixed-initiative interactions, meaning that the system should proactively ask relevant questions for eliciting preferences and clarifying information needs [16]. My work was dedicated to advancing our understanding and highlighting differences in patterns of such mixed-initiative interactions between human-machine and human-human information-seeking dialogues. I also made a methodological contribution by introducing a novel training approach that allows language models to generate more diverse questions and demonstrated its effectiveness in practice.

My work presented here contributes novel approaches, findings and valuable datasets that were made publicly available to the research community. These machine-learning datasets constitute important resources being actively used for model training and evaluation. While the majority of research described in this habilitation thesis preceded the arrival of LLMs, it was applied to the previous generation of language models that use the same machine-learning architecture and similar training approaches. This makes the main findings either directly transferable to LLMs or clearly point towards analogous patterns expected in larger models.

The publications that constitute this cumulative thesis has led to the following major research outputs: (i) the first large-scale dataset for conversational QA over the Web; (ii) the novel approach of question rewriting that allows an LLM to interact with a search engine to answer conversational questions; (iii) the first large-scale analysis of patterns of mixed initiative in information-seeking dialogues that revealed the discrepancies between the human-machine and human-human dialogues; (iv) the first automatically constructed dataset for the entity linking task that focuses on the ability of language models to correctly interpret context and disambiguate entity mentions; (v) the first automatically constructed dataset of public court opinions that focuses on the ability of language models to summarize legal documents and propose legal citations.

2. Summary of the Publications

This habilitation thesis comprises a collection of articles relating to the topic of conversational search including question answering, retrieval-augmented generation, summarization, question rewriting, as well as an extension to the analysis of mixed initiative in dialogues and question generation.

Question Rewriting. This line of work that was initiated due to my internship at Apple Seattle in 2019 directly led to my two top-cited publications, achieving, in total, over 400 citations since only 2021, which can be considered as "groundbreaking achievement" (cf. habilitation guidelines).

Paper (1) entitled "Question Rewriting for Conversational Question Answering" [41] was the first to introduce the novel approach of question rewriting that employs an LM to reformulate the conversational questions, such that they can be efficiently answered outside of the conversation context. This work is an outcome of the collaboration with researchers at Apple Seattle during and after my internship there. In this paper, we use previously introduced datasets for the task of open-domain conversational QA. We also made an extensive error analysis and released a software toolkit that allows to automatically differentiate between different error types: stemming from the question rewriting vs QA errors. The resulting follow-up paper [40] received the best paper award at the SCAI workshop co-located with EMNLP in 2020.

Paper (2) entitled "Question Answering Goes Conversational via Question Rewriting" [3] builds on paper (1) and extends it with the new dataset, QReCC, specifically designed for the task of open-domain conversational QA with question rewriting. This dataset was collected during several months and is the result of a collaborative work with an internal team of professional annotators at Apple, who were instructed to come up with plausible conversational questions one might ask a voice assistant, such as Siri. The annotators were also tasked to find answers to these conversational questions on the Web, formulate them in the context of the conversation and provide the URLs to the sources used to produce those answers. Moreover, in this paper, we also introduced an extension to the Transformer architecture that effectively boosted performance on the question rewriting task. The dataset of conversational question-answers that we introduced in this work is subsequently being used for training LLMs. In particular, it was used by Google to fine-tune Flan models, which became the first widely-used multi-task language models [11]. The Hugging Face repository currently lists 65 open models that were fine-tuned on our QReCC dataset.

Paper (3) entitled "SCAI-QReCC Shared Task on Conversational Question Answering" [39] was a direct follow-up on the previous two papers. In this task, the original QReCC dataset was further extended with alternative generated answers that were submitted by different teams and evaluated by the crowd workers. This shared task was part of the SCAI 2021 workshop and included both question rewriting and answer generation subtasks.

Paper (4) entitled "Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models" [45] is the first publication of the PhD student jointly co-supervised with Prof. Kanoulas at the University of Amsterdam. While the work on this project was started already back in 2020, it was not possible to publish it for a long time due to the fierce competition that broke out in the research community developing LLMs and RAG approaches in parallel. This project set out to explore the connection between RAG and query-focused summarization by establishing the link between the two tasks and evaluating modern RAG approaches against more traditional summarization approaches in a new setting where the summarization task was extended to a multi-document open-domain scenario. This subtle historical connection was largely

overlooked in the modern day LLM-heavy research and my hope is that this paper may bring back some of the attention to previous work on QA and summarization, which it deserves.

Paper (5) entitled "Neural Ranking with Weak Supervision for Open-Domain Question Answering: A Survey" [33] was conducted in collaboration with my new colleagues when I started at Amazon Alexa AI. The motivation behind this paper is the fact that one of the critical barriers that hinder further improvement of the RAG models is the absence of training data. Therefore, we set out to explore different approaches and techniques that were introduced in the recent years and were designed to circumvent this issue to some extent. We came up with a new framework that allowed to better group, describe and understand these approaches and relations between them.

Mixed Initiative. This line of work extends and diversifies my research contributions beyond the popular task of QA towards a more comprehensive mixed-initiative human-machine dialogue.

Paper (6) entitled "A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search" [38] is a journal paper published at ACM TOIS during my PostDoc research at the University of Amsterdam. It extends the conversational analysis approach that I introduced first in my PhD thesis and extends it with the analysis for mixed initiative patterns in dialogues that we first introduced in the short ACM SIGIR paper in 2020 [37].

Paper (7) entitled "Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation" [35] is my first-author research publication that appeared in ECAI and was based on the research I have conducted at Amazon Alexa AI. The main motivation for this project was the need to generate follow-up suggested questions in a conversational AI interface. In this paper, I proposed a novel approach designed to make question generation more controllable and diverse.

Benchmarking. Those two papers were co-authored during my PostDoc research at the University of Amsterdam, when I co-supervised several PhD students with Prof. Kanoulas. Both of those publications study the fundamental properties of language modeling technology to identify research gaps and track progress on complex tasks that require accurate natural language understanding.

Paper (8) entitled "Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing" [24] demonstrated the shortcomings of the state-of-the-art language models, which exhibit frequency bias since they are explicitly trained to predict next word probability distribution based on word co-occurrences in their training data. To this end, we conducted an experimental evaluation that revealed how often language models fail to correctly incorporate context when disambiguating between multiple alternative interpretations defaulting to the most frequent one regardless of the provided context.

Paper (9) entitled "VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law" [29] showed that summarizing long legal texts remain challenging. We collected publicly available court opinions, automatically constructed a citation graph and evaluated whether language models can be used to propose legal citations.

Further, complementary contributions. I had an honor to work with many PhD, master and bachelor students from the University of Amsterdam. These collaborations often led to the joint publications, in which I played the role of a co-supervisor, and which I am also very proud of:

- I consistently supported the work of Dr. Mariya Hendriksen, who focused on image retrieval, and which led to two full paper publications in ECIR (CORE A ranking) [12, 13] that contributed to her successful PhD defense at the University of Amsterdam in 2024.
- I was mentoring Dr. Philipp Christmann during his internship at Amazon Artificial General Intelligence (AGI) in 2023-2024. The outcomes of this research project were published in Findings of EMNLP 2024 (CORE A* ranking) as a short paper [10] (acceptance rate for Findings papers at EMNLP 2024 was 16.9% [1]).
- Dr. Georgios Sidiropoulos was a PhD student co-supervised with Prof. Kanoulas, who published a short paper at CIKM [34] (CORE A ranking) based on my initial experiments with spoken QA.

- M. Eng. Shaojie Jiang is a PhD student co-supervised with Prof. de Rijke and expected to graduate in 2026. His work on X was published in CHIIR [14], which is one of the prime venues for IR research (SJR of 0.178 and an h-index of 13).

- BSc Niels Rouws worked on QA for Dutch language under joint supervision by Dr. Katrenko and me towards his BSc thesis at the University of Amsterdam leading to his first research publication [30].

In addition to that, I remain the core organizer of the SCAI workshop since 2021 through 2025 [36, 23], which is an established venue bringing researchers working on various subtasks related to conversational search, such as IR, natural language generation and evaluation. I also co-organised the DialDoc workshop at ACL 2022 and the TREC Conversational Assistance Track (CAsT) in 2022.

This work lead me to the new research questions and a follow-up research agenda that has recently received generous founding from Vienna Science and Technology Fund (WWTF) under the Vienna Research Group (VRG) grant (IA 27002141).

VRG Research Proposal: Knowledge Representation Learning for LLMs. Our previous research shows that RAG works well in some cases, such as for answering simple factoid questions, if the relevant information is at the top of the search results and the LLM is fine-tuned to use them when generating the response. However, in practice it fails in many cases when information needs are more complex and search results are not optimal. In this proposal, I advocate for the need to address the aforementioned problem upfront: instead of "plugging in" an existing IR system, a new solution specifically designed with an LLM in mind should be conceived. While embeddings-based approaches, such as dense retrieval were proposed to this end, they still rely on the same patterns from classic IR systems, such as inverted index, and do not address the fundamental problem of knowledge representation. Alternative symbolic approaches for modeling knowledge, such as knowledge graphs, also follow strict structural rules that define what and how information should be represented. The main novelty of my proposal is to give an LLM the tools necessary to find the optimal structure for its internal symbolic knowledge representation by itself, and that can best support and align with its internal subsymbolic knowledge of the language accumulated in its parameters. To accomplish this, an objective function shall be defined in terms of the success criteria for the resulting knowledge representation, in terms of its efficacy in supporting faithful response generation and structural properties based on the database theory that are more likely to efficiently guide the model towards the global optimum. Designing an internal knowledge representation for an LLM is the major goal of the project. In addition, my aim is to ensure that the proposed architecture has a clear potential to scale up to encorporate all of the Web's knowledge, which can only be feasible when embracing the decentralised nature of the Web as a network of interconnected data repositories.

Importantly, embedding this research project within WU Wien's ecosystem will also allow us to actively collaborate with legal researchers: although the regulation of AI technology is already well under way across the EU states, as manifested in the recent EU AI Act and the General Data Protection Regulation (GDPR), there is a clear need for a more active collaboration between AI researchers and legal experts to ensure that the new architectures are designed in compliance with legal requirements and to put mechanisms in place that allow to examine the system and enforce these requirements. It is not possible to halt the advance of the modern technologies, such as LLMs, at this stage, since they clearly bring important benefits for the modern society but it is also of great importance to be aware of shortcomings and limitations that can lead to undesirable outcomes. It is important to ensure that further development of this technology is geared towards making it more safe, fair and transparent for the end users, businesses and legislators putting them also in a position to co-create the future of modern AI technologies rather than forcing them to adapt to it post factum.

3. Fulfilment of the Requirements

The habilitation guide of the Department of Information Systems and Operations at the Vienna University of Economics and Business stipulates that a cumulative habilitation thesis should consist of at least five excellent academic articles that are thematically related with each article offering distinctive contributions. The number of articles indicated above serves as a reference point and may be reduced if the habilitation candidate has produced and published groundbreaking scientific achievements. The formal requirements are assessed as follows:

1. **Excellence in research**. In order to ensure scientific excellence and to attest to the novelty and expected scientific impact of the work, these articles should have gone through a rigorous review process. The articles are expected to be published in well-known and highly ranked scientific journals or excellent conference proceedings. The externally validated CORE rating can serve as evidence for the quality of the conference proceedings. A maximum of two *excellent* academic articles (CORE A* rating) can be substituted by three *very good* conference contributions (CORE A and B ratings) per article, under the assumption that the respective contributions are published as full papers, the conferences are organized by major discipline-specific associations that have a rigorous review processes (e.g. ACM, IEEE, IFIP, Usenix, AIS) and an acceptance rate of less than 30%. If an article has several authors, the habilitation candidate is expected to clearly articulate their specific contribution to the article.

2. **Demonstration of scientific independence**. In order to demonstrate the candidate's lead role in conducting research, at least one of the following criteria should apply: either (a) one of the articles must be a single-author publication; or (b) at least three first-author publications; or (c) a successful acquisition of competitive research funding as a principal investigator in a single-applicant or ad personal program (e.g., FWF, ERC, WWTF), whose funding is sufficient to support at least a postdoc or a PhD position.

The candidate's cumulative habilitation thesis fulfills both of the aforementioned requirements. In short, the thesis is composed of nine high-quality peer-reviewed publications with the majority (five) of them being first-author publications, including one of them published in the top-ranked journal, and the rest are excellent and very good conference publications in prestigious Computer Science research venues. In all cases, the Contributor Roles Taxonomy (CRediT) is used to clearly articulate the candidate's contribution to these publications. The metrics below were taken from the Scimago, Resurchify and the official CORE ranking:

- ACM Transactions on Information Systems (TOIS), ACM, Quartile Q1, h-index=100, Impact Score=9.17, SJR=1.54, JCR=9.1 (10th out of 258 in Computer Science and Information Systems).

- ACM International Conference on Web Search and Data Mining (WSDM), ACM, Field Of Research: 4611 - Machine learning, Field Of Research: 4605 - Data management and data science, CORE2021 Rank: A* = Exceptional (flagship conference), 7.64% of 785 ranked venues, Acceptance rate: 18.86% (Most Recent Year 2013).

- Empirical Methods in Natural Language Processing (EMNLP), ACL, Field Of Research: 4602 - Artificial intelligence, CORE2023 Rank: A* = Exceptional (flagship conference), 7.64% of 785 ranked venues, Acceptance rate: 22% (Most Recent Year 2022).
- European Conference on Artificial Intelligence (ECAI), EurAI, Field Of Research: 4602 - Artificial intelligence, Field Of Research: 4611 - Machine learning, CORE2023 Rank: A = Excellent, 14.9% of 785 ranked venues, Acceptance rate: 20% (Most Recent Year 2018).
- ACM International Conference on Information and Knowledge Management (CIKM), ACM, Field Of Research: 4602 - Artificial intelligence, 4605 - Data management and data science, CORE2023 Rank: A = Excellent, 14.9% of 785 ranked venues, Acceptance rate: 20% (Most Recent Year 2019).
- North American Association for Computational Linguistics (NAACL), ACL, Field Of Research: 4602 - Artificial intelligence, CORE2023 Rank: A = Excellent, 14.9% of 785 ranked venues, Acceptance rate: 21% (Most Recent Year 2022).
- European Association for Computational Linguistics (EACL), ACL, Field Of Research: 4602 - Artificial intelligence, CORE2023 Rank: A = Excellent, 14.9% of 785 ranked venues, Acceptance rate: 24% (Most Recent Year 2023).
- International Conference on Pattern Recognition (ICPR), Field Of Research: 4603 - Computer vision and multimedia computation, CORE2023 Rank: B - 28.15% of 785 ranked venues, Acceptance rate: 52% (Most Recent Year 2018).
- Language Resources and Evaluation Conference (LREC), ELRA, Field Of Research: 4602 - Artificial intelligence, CORE2023 Rank: B - 28.15% of 785 ranked venues, Acceptance rate: 62% (Most Recent Year 2022).

In the following, we use the CRediT roles in order to clarify the habilitation candidates contribution to the articles that comprise this cumulative habilitation thesis. Additionally, we provide a short justification for inclusion of each article.

1. **Svitlana Vakulenko**, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: Question Rewriting for Conversational Question Answering. WSDM 2021: 355-363.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization, and Project administration.

   Justification for inclusion: This is an A* highly-cited conference contribution (Acceptance rate: 18.86%, citations at the time of submission: 224 since 2021), where the candidate is the lead author. She introduced and motivated the idea of question rewriting, conducted the initial experiments with question rewriting during her internship in Apple, proposed the experimental setup and was responsible for co-ordinating the evaluation process among the other three core team members at Apple, and performed the qualitative analysis as well as the majority of the paper write-up.

2. Raviteja Anantha*, **Svitlana Vakulenko***, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, and Project administration.

Justification for inclusion: This is an A highly-cited conference contribution (citations at the time of submission: 228 since 2021), where the candidate is one of the two lead authors with equal contributions indicated in the head of the paper (the other lead author was my internship supervisor at Apple). The candidate provided detailed annotation instructions and supervised the initial data collection process at Apple that was scaled later on, proposed the experimental setup and was responsible for co-ordinating the evaluation process among the other three core team members at Apple, performed the qualitative analysis and contributed the major part of the paper write-up.

3. **Svitlana Vakulenko**, Johannes Kiesel, Maik Fröbe: SCAI-QReCC Shared Task on Conversational Question Answering. LREC 2022: 4913-4922.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, and Project administration.

   Justification for inclusion: The habilitation candidate is the first and main author of this article. She was responsible for co-ordinating organisation of the shared task, setting up the annotation task for evaluating the submissions and writing up the report.

4. Weijia Zhang, Jia-Hong Huang, **Svitlana Vakulenko**, Yumo Xu, Thilina Rajapakse, Evangelos Kanoulas: Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models. ICPR (19) 2024: 89-104.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Writing - Review & Editing, and Supervision.

   Justification for inclusion: The habilitation candidate was one of two supervisors of this work. She proposed the initial idea, helped the main author to prepare the PhD proposal and start the work on this project.

5. Xiaoyu Shen, **Svitlana Vakulenko**, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, Adrià de Gispert: Neural Ranking with Weak Supervision for Open-Domain Question Answering : A Survey. EACL (Findings) 2023: 1691-1705.

   Habilitation Candidate CRediT roles: Conceptualisation, Writing - Review & Editing, Visualization.

   Justification for inclusion: The habilitation candidate contributed her expertise in open-domain QA to this article. She helped to conceptualize the framework and designed the core visualization supporting the structure of the survey.

6. **Svitlana Vakulenko**, Evangelos Kanoulas, Maarten de Rijke: A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. ACM Trans. Inf. Syst. 39(4): 49:1-49:32 (2021).

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft, and Writing - Review & Editing, Visualization.

   Justification for inclusion: The habilitation candidate is the first and main author of this article. She collected the datasets, conducted the analysis and prepared the initial draft for the TOIS special issue on conversational search. She has further refined the draft based on the reviewers' comments and carried out the submission process until the final acceptance decision.

7. **Svitlana Vakulenko**, Bill Byrne, Adrià de Gispert: Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation. ECAI 2023: 2378-2385.

Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Software, Validation, Investigation, Data Curation, Writing - Original Draft, and Writing - Review & Editing.

Justification for inclusion: The habilitation candidate is the first and main author of this article. She proposed the idea, prepared the datasets, organized and guided professional annotation tasks, implemented modifications to the training procedure, carried out and documented the experiments.

8. Vera Provatorova, Samarth Bhargav, **Svitlana Vakulenko**, Evangelos Kanoulas: Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing. EMNLP (1) 2021: 10501-10510.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Writing - Review & Editing, and Supervision.

   Justification for inclusion: The habilitation candidate was one of two supervisors of this work. She contributed the main ideas for this project and closely supervised the main author when preparing and conducting the experiments.

9. Julien Rossi, **Svitlana Vakulenko**, Evangelos Kanoulas: VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law. CIKM 2021: 4554-4563.

   Habilitation Candidate CRediT roles: Conceptualisation, Methodology, Writing - Review & Editing, and Supervision.

   Justification for inclusion: The habilitation candidate was one of two supervisors of this work. She contributed ideas on improving the publication draft, strengthen the experimental evaluation and helped to prepare the submission.

Last but not least, scientific independence of the habilitation candidate has been demonstrated by the successful acquisition of competitive research funding from WWTF, which amounts to the total of more than 1.5M Euros and supports filling three PhD positions. The candidate already took on the role of the principal investigator on this project and the leader of the newly established Vienna Research Group (VRG) on Knowledge Representation Learning for LLMs, which is also the first VRG funded at WU.

Bibliography

[1] AL-ONAIZAN, Y., BANSAL, M., AND CHEN, Y.-N. Findings of the association for computational linguistics: Emnlp 2024. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (2024).

[2] ANAND, A., CAVEDON, L., HAGEN, M., JOHO, H., SANDERSON, M., AND STEIN, B. Conversational search: Dagstuhl seminar 19461 report. *Dagstuhl Reports 9*, 11 (2020), 34–83.

[3] ANANTHA, R., VAKULENKO, S., TU, Z., LONGPRE, S., PULMAN, S., AND CHAPPIDI, S. Open-domain question answering goes conversational via question rewriting. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)* (2021), K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds., Association for Computational Linguistics, pp. 520–534.

[4] BAEZA-YATES, R., AND RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison-Wesley, Boston, MA, 1999.

[5] BATES, M. J. The design of browsing and berry-picking techniques for the online search interface. In *Online Review* (1989), vol. 13, pp. 407–424. Introduces the Berry Picking model, emphasizing iterative, evolving search behavior in interactive information retrieval.

[6] BELKIN, N. J., ODDY, R. N., AND BROOKS, H. M. Ask for information retrieval: Part i. background and theory. *Journal of Documentation 38*, 2 (1982), 61–71. Introduces the Anomalous States of Knowledge (ASK) model, emphasizing interactive IR and iterative query refinement.

[7] BURTSEV, M., CHUKLIN, A., KISELEVA, J., AND BORISOV, A. Search-Oriented Conversational AI (SCAI). In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017* (2017), J. Kamps, E. Kanoulas, M. de Rijke, H. Fang, and E. Yilmaz, Eds., ACM, pp. 333–334.

[8] BUSH, V. As we may think. *The Atlantic Monthly 176*, 1 (July 1945), 101–108. Available online: `https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/`.

[9] BÜTTCHER, S., CLARKE, C. L. A., AND CORMACK, G. V. *Information Retrieval: Implementing and Evaluating Search Engines*, 1st ed. The MIT Press, Cambridge, MA, 2010.

[10] CHRISTMANN, P., VAKULENKO, S., SORODOC, I., BYRNE, B., AND DE GISPERT, A. Retrieving contextual information for long-form question answering using weak supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024* (2024), Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds., Association for Computational Linguistics, pp. 14301–14310.

[11] CHUNG, H. W., HOU, L., LONGPRE, S., ZOPH, B., TAY, Y., FEDUS, W., LI, Y., WANG, X., DEHGHANI, M., BRAHMA, S., WEBSON, A., GU, S. S., DAI, Z., SUZGUN, M., CHEN, X., CHOWDHERY, A., CASTRO-ROS, A., PELLAT, M., ROBINSON, K., VALTER, D., NARANG, S., MISHRA, G., YU, A., ZHAO, V. Y., HUANG, Y., DAI, A. M., YU, H., PETROV, S., CHI, E. H., DEAN, J., DEVLIN, J., ROBERTS, A., ZHOU, D., LE, Q. V., AND WEI, J. Scaling instruction-finetuned language models. *J. Mach. Learn. Res. 25* (2024), 70:1–70:53.

[12] HENDRIKSEN, M., BLEEKER, M. J. R., VAKULENKO, S., VAN NOORD, N., KUIPER, E., AND DE RIJKE, M. Extending CLIP for category-to-image retrieval in e-commerce. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I* (2022), M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, and V. Setty, Eds., vol. 13185 of *Lecture Notes in Computer Science*, Springer, pp. 289–303.

[13] HENDRIKSEN, M., VAKULENKO, S., KUIPER, E., AND DE RIJKE, M. Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, Eds., vol. 13982 of *Lecture Notes in Computer Science*, pp. 68–85.

[14] JIANG, S., VAKULENKO, S., AND DE RIJKE, M. Weakly supervised turn-level engagingness evaluator for dialogues. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, March 19-23, 2023* (2023), J. Gwizdka and S. Y. Rieh, Eds., ACM, pp. 258–268.

[15] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W., ROCKTÄSCHEL, T., RIEDEL, S., AND KIELA, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds.

[16] LU, L., MENG, C., RAVENDA, F., ALIANNEJADI, M., AND CRESTANI, F. Zero-shot and efficient clarification need prediction in conversational search. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part I* (2025), C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, Eds., vol. 15572 of *Lecture Notes in Computer Science*, Springer, pp. 389–404.

[17] LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development 1*, 4 (1957), 309–317. Early work on automatic indexing; introduced term-to-document mapping concepts precursor to inverted indexes.

[18] LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development 2*, 2 (1959), 159–165. Introduced KWIC indexing; foundational work for inverted index structures.

[19] MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[20] MINAEE, S., MIKOLOV, T., NIKZAD, N., CHENAGHLU, M., SOCHER, R., ET AL. Large language models: A survey. *arXiv preprint* (2024).

[21] MO, F., MENG, C., ALIANNEJADI, M., AND NIE, J. Conversational search: From fundamentals to frontiers in the llm era. Tutorial at the 48th ACM SIGIR Conference on Research and Development in Information Retrieval, 2025. Padua, Italy. Available as arXiv preprint.

[22] MOOERS, C. N. Zator technical bulletin no. 48. Tech. rep., Zator Company, 1950. Contains Mooers' early definition of "information retrieval," including the phrase: "The requirements of information retrieval, of finding information whose location or very existence is a priori unknown.".

[23] PENHA, G., VAKULENKO, S., DUSEK, O., CLARK, L., PAL, V., AND ADLAKHA, V. The seventh workshop on search-oriented conversational artificial intelligence (scai'22). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022* (2022), E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds., ACM, pp. 3466–3469.

[24] PROVATOROVA, V., BHARGAV, S., VAKULENKO, S., AND KANOULAS, E. Robustness evaluation of entity disambiguation using prior probes: the case of entity overshadowing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021* (2021), M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Association for Computational Linguistics, pp. 10501–10510.

[25] RADLINSKI, F., AND CRASWELL, N. A theoretical framework for conversational search. In *Proceedings of the ACM Conference on Human Information Interaction and Retrieval (CHIIR)* (2017), ACM, pp. 117–126. Introduces a theoretical model for conversational search systems and defines desirable properties for conversational IR.

[26] REDDY, S., CHEN, D., AND MANNING, C. D. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics 7* (2019), 249–266.

[27] ROBERTSON, S. E., AND JONES, K. S. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 3 (1976), 129–146. Foundational paper on probabilistic models of information retrieval; precursor to BM25.

[28] ROBERTSON, S. E., WALKER, S., JONES, S., HANCOCK-BEAULIEU, M., AND GATFORD, M. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)* (1995), National Institute of Standards and Technology (NIST), pp. 109–126. Introduces BM25 scoring function for probabilistic information retrieval.

[29] ROSSI, J., VAKULENKO, S., AND KANOULAS, E. Verbcl: A dataset of verbatim quotes for highlight extraction in case law. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021* (2021), G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, and H. Tong, Eds., ACM, pp. 4554–4563.

[30] ROUWS, N. J., VAKULENKO, S., AND KATRENKO, S. Dutch squad and ensemble learning for question answering from labour agreements. In *Artificial Intelligence and Machine Learning - 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10-12, 2021, Revised Selected Papers* (2021), L. A. Leiva, C. Pruski, R. Markovich, A. Najjar, and C. Schommer, Eds., vol. 1530 of *Communications in Computer and Information Science*, Springer, pp. 155–169.

[31] SALTON, G. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York, 1968. Describes inverted file structures and early automatic indexing techniques; foundational for IR.

[32] SALTON, G., WONG, A., AND YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM 18*, 11 (1975), 613–620. Formalizes inverted index and introduces the vector space model for document ranking.

[33] SHEN, X., VAKULENKO, S., TREDICI, M. D., BARLACCHI, G., BYRNE, B., AND DE GISPERT, A. Neural ranking with weak supervision for open-domain question answering : A survey. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023* (2023), A. Vlachos and I. Augenstein, Eds., Association for Computational Linguistics, pp. 1691–1705.

[34] SIDIROPOULOS, G., VAKULENKO, S., AND KANOULAS, E. On the impact of speech recognition errors in passage retrieval for spoken question answering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022* (2022), M. A. Hasan and L. Xiong, Eds., ACM, pp. 4485–4489.

[35] VAKULENKO, S., BYRNE, B., AND DE GISPERT, A. Uniform training and marginal decoding for multi-reference question-answer generation. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)* (2023), K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, and R. Radulescu, Eds., vol. 372 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 2378–2385.

[36] VAKULENKO, S., DUSEK, O., AND CLARK, L. Report on the 6th workshop on search-oriented conversational AI (SCAI 2021). *SIGIR Forum 55*, 2 (2021), 20:1–20:14.

[37] VAKULENKO, S., KANOULAS, E., AND DE RIJKE, M. An analysis of mixed initiative and collaboration in information-seeking dialogues. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020* (2020), J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds., ACM, pp. 2085–2088.

[38] VAKULENKO, S., KANOULAS, E., AND DE RIJKE, M. A large-scale analysis of mixed initiative in information-seeking dialogues for conversational search. *ACM Trans. Inf. Syst. 39*, 4 (2021), 49:1–49:32.

[39] VAKULENKO, S., KIESEL, J., AND FRÖBE, M. Scai-qrecc shared task on conversational question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022* (2022), N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., European Language Resources Association, pp. 4913–4922.

[40] VAKULENKO, S., LONGPRE, S., TU, Z., AND ANANTHA, R. A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)* (2020), pp. 7–16.

[41] VAKULENKO, S., LONGPRE, S., TU, Z., AND ANANTHA, R. Question rewriting for conversational question answering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021* (2021), L. Lewin-Eytan, D. Carmel, E. Yom-Tov, E. Agichtein, and E. Gabrilovich, Eds., ACM, pp. 355–363.

[42] VAN RIJSBERGEN, C. J. *Information Retrieval*. Butterworths, London, UK, 1979.

[43] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), pp. 5998–6008.

[44] WALKER, M. A., SIDNER, C. L., AND LITMAN, S. M. Learner: A multi-turn spoken dialogue system for information access. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using NLP* (2003), pp. 25–32. Describes a multi-turn spoken dialogue system for interactive information access, an early contribution to dialogue-based IR.

[45] ZHANG, W., HUANG, J., VAKULENKO, S., XU, Y., RAJAPAKSE, T., AND KANOULAS, E. Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In *Pattern Recognition - 27th International Conference, ICPR 2024, Kolkata, India, December 1-5, 2024, Proceedings, Part XIX* (2024), A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C. Liu, S. Bhattacharya, and U. Pal, Eds., vol. 15319 of *Lecture Notes in Computer Science*, Springer, pp. 89–104.

Part II

Publications

1. Question Rewriting for Conversational Question Answering

Bibliographic Information

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, Raviteja Anantha: Question Rewriting for Conversational Question Answering. WSDM 2021: 355-363.

Question Rewriting for Conversational Question Answering

Svitlana Vakulenko*
University of Amsterdam
s.vakulenko@uva.nl

Shayne Longpre
Apple Inc.
slongpre@apple.com

Zhucheng Tu
Apple Inc.
zhucheng_tu@apple.com

Raviteja Anantha
Apple Inc.
raviteja_anantha@apple.com

ABSTRACT

Conversational question answering (QA) requires the ability to correctly interpret a question in the context of previous conversation turns. We address the conversational QA task by decomposing it into question rewriting and question answering subtasks. The question rewriting (QR) subtask is specifically designed to reformulate ambiguous questions, which depend on the conversational context, into unambiguous questions that can be correctly interpreted outside of the conversational context. We introduce a conversational QA architecture that sets the new state of the art on the TREC CAsT 2019 passage retrieval dataset. Moreover, we show that the same QR model improves QA performance on the QuAC dataset with respect to answer span extraction, which is the next step in QA after passage retrieval. Our evaluation results indicate that the QR model we proposed achieves near human-level performance on both datasets and the gap in performance on the end-to-end conversational QA task is attributed mostly to the errors in QA.

CCS CONCEPTS

• **Information systems** **Question answering**; **Query reformulation**; **Information retrieval**; **Information extraction**.

KEYWORDS

conversational search, question answering, question rewriting

ACM Reference Format:
Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. Question Rewriting for Conversational Question Answering. In *preparation for ACM Conference on Web Search and Data Mining, March 8–12, 2021, Jerusalem, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Extending question answering systems to a conversational setting is an important development towards a more natural human-computer interaction [9]. In this setting a user is able to ask several follow-up questions, which omit but reference information that was already introduced earlier in the same conversation, for example:

(Q) - Where is Xi'an?
(A) - Shaanxi, China.
(Q) - What is **its** GDP?
(A) - 95 Billion USD.
(Q) - What is the share *(of Xi'an)* in the *(Shaanxi)* province GDP?
(A) - 41.8% of Shaanxi's total GDP

This example highlights two linguistic phenomena characteristic of a human dialogue, which include **anaphora** (words that explicitly reference previous conversation turns) and *ellipsis* (words that can be omitted from the conversation) [35]. Therefore, conversational QA models require mechanisms capable of resolving contextual dependencies to correctly interpret such follow-up questions. While existing co-reference resolution tools are designed to handle anaphoras, they do not provide any support for ellipsis [4]. Previous research showed that QA models can be directly extended to incorporate conversation history but a considerable room for improvement still remains especially when scaling such models beyond a single input document [2, 26, 31].

In this paper we show how to extend existing state-of-the-art QA models with a QR component and demonstrate that the proposed approach improves the performance on the end-to-end conversational QA task. This QR component is specifically designed to handle ambiguity of the follow-up questions by rewriting them such that they can be processed by existing QA models as stand-alone questions outside of the conversation context. This setup also offers a wide range of practical advantages over the single end-to-end conversational QA model:

- **Traceability**: QR component produces the question that the QA model is actually trying to answer, which allows to separate errors that stem from an incorrect question interpretation from errors in question answering. Our error analysis makes full use of this feature by investigating different sources of errors and their correlation.

- **Reuse**: QR allows to reduce the conversational QA task to the standard QA task and leverage already existing QA models and datasets. Any new non-conversational QA model can be immediately ported into a conversational setting using QR. A single QR model can be reused across several alternative QA architectures as we show in our experiments.

- **Modularity**: In practice, information is distributed across a network of heterogeneous nodes that do not share internal

*This research was completed during the internship at Apple Inc.

WSDM '21, March 8–12, 2021, Jerusalem, Israel

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

representations. A natural language question can provide an adequate communication protocol between these distributed components [30]. Instead of sending the content of the whole conversation to a 3rd-party API, QR allows for formulating a concise question that contains only information relevant to the current information need, which also helps to determine which of the distributed systems should be used for answering the question.

Two dimensions on which QA tasks vary are: the type of data source used to retrieve the answer (e.g., a paragraph, a document collection, or a knowledge graph); and the expected answer type (a text span, a ranked list of passages, or an entity). In this paper, we experiment with two variants of the QA task: *retrieval QA*, the task of finding an answer to a given natural-language question as a ranked list of relevant passages given a document collection; and *extractive QA*, the task of finding an answer to a given natural-language question as a text span within a given passage. Though the two QA tasks are complementary to each other, in this paper we focus on the QR task and its ability to enable different types of QA models within a conversational setting. We experiment with both retrieval and extractive QA models to examine the effect of the QR component on the end-to-end QA performance. The contributions of this work are three-fold:

(1) We introduce a novel approach for the conversational QA task based on question rewriting that sets the new state-of-the-art results on the TREC CAsT dataset for passage retrieval.
(2) We show that the same question rewriting model (trained on the same dataset) boosts the performance of the state-of-the-art architecture for answer extraction on the QuAC dataset.
(3) We systematically evaluate the proposed approach in both retrieval and extractive settings, and report the results of our error analysis.

2 RELATED WORK

Conversational QA is an extension of the standard QA task that introduces contextual dependencies between the input question and the previous dialogue turns. Several datasets were recently proposed extending different QA tasks to a conversational setting including extractive [2, 31], retrieval [4] and knowledge graph QA [3, 12]. One common approach to conversational QA is to extend the input of a QA model by appending previous conversation turns [3, 14, 27]. Such approach, however falls short in case of retrieval QA, which requires a concise query as input to the candidate selection step, such as BM25 [24]. Results of the recent TREC CAsT track demonstrated that co-reference models are also not sufficient to resolve the missing context in the follow-up questions [4]. A considerable gap between the performance of automated rewriting approaches and manual human annotations call for new architectures that are capable of retrieving relevant answers from large text collections using conversational context.

Query rewriting is a technique that was successfully applied in a variety of tasks, including data integration, query optimization and query expansion in sponsored search [11, 25, 33]. More recently rewriting conditioned on the conversation history was shown to increase effectiveness of chitchat and task-oriented dialogue models [30, 34]. QR in the context of the conversational QA task was first introduced by Elgohary et al. [7], who released the CANARD dataset that contains human rewrites of questions from QuAC conversational reading comprehension dataset. Their evaluation, however, was limited to analysing rewriting quality without assessing impact from rewriting on the end-to-end conversational QA task. Our study was designed to bridge this gap and extend evaluation of question rewriting for conversational QA to the passage retrieval task as well.

The field of conversational QA is advancing rapidly due to the challenges organised by the community, such as the TREC CAsT challenge [4]. Several recent studies that are concurrent to our work proposed alternative approaches for the conversational QA task and evaluated them on the TREC CAsT dataset. Mele et al. [21] introduced an unsupervised approach that relies on a set of heuristics using part-of-speech tags, dependency parses and co-reference resolution to rewrite questions. Voskarides et al. [37] train a binary classification model on CANARD dataset that learns to pick the terms from the conversation history for query expansion. Yu et al. [41] train a sequence generation model using a set of rules applied to the MS MARCO search sessions. We demonstrate that our approach does not only outperform all the results reported on the TREC CAsT dataset previously but also boosts the performance on the answer extraction task as demonstrated on the QuAC dataset.

3 QUESTION REWRITING TASK

We assume the standard setting of a conversational (sequential) QA task, in which a user asks a sequence of questions and the system is to provide an answer after each of the questions. Every follow-up question may be ambiguous, i.e., its interpretation depends on the previous conversation turns (example: "What is its GDP?"). The task of question rewriting is to translate every ambiguous follow-up question into a semantically equivalent but unambigous question (for this example: "What is the GDP of Xi'an?"). Then, every question is first processed by the QR component before passing it to the QA component.

More formally, given a conversation context C and a potentially implicit question Q, a question which may require the conversation context C to be fully interpretable, the task of a question rewriting (QR) model is to generate an explicit question Q' which is equivalent to Q under conversation context C and has the same correct answer A. See Figure 1 for an illustrative example that shows how QR can be combined with any non-conversational QA model.

4 APPROACH

In this section, we describe the model architectures of our QR and QA components. We use two distinct QA architectures to show that the same QR model can be successfully applied across several QA models irrespective of their implementation. One of the QA models is designed for the task of passage retrieval and the other one for the task of answer extraction from a passage (reading comprehensions). We employ state-of-the-art QA architectures that were already successfully evaluated for these tasks in non-conversational settings and show how QR allows to port existing QA models into a conversational setting.

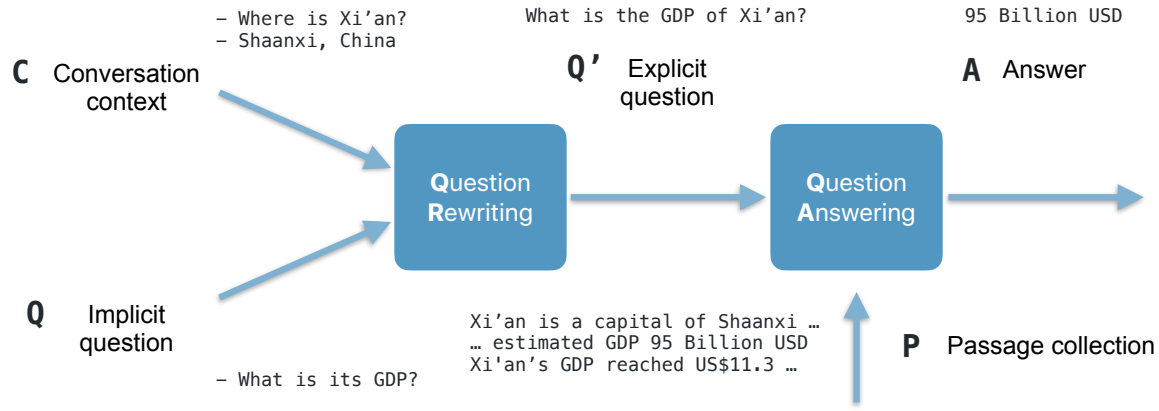


Figure 1: Our approach for end-to-end conversational QA relies on the question rewriting component to handle conversation context and produce an explicit question that can be fed to standard, non-conversational QA components.

4.1 Question Rewriting Model

We use a model for question rewriting, which employs a unidirectional Transformer decoder [28] for both encoding the input sequence and decoding the output sequence. The input to the model is the question with previous conversation turns (we use 5 previous turns in our experiments) turned into token sequences separated with a special [*SEP*] token.

The training objective is to predict the output tokens provided in the ground truth question rewrites produced by human annotators. The model is trained via teacher forcing approach, which is a standard technique for training language generation models, to predict every next token in the output sequence given all the preceding tokens. The loss is calculated via negative log-likelihood (cross-entropy) between the output distribution $D' \in \mathbb{R}^{|V|}$ over the vocabulary V, and the one-hot vector $y_{Q'} \in \mathbb{R}^{|V|}$ for the correct token from the ground truth: $loss = -y_{Q'} \log D'$.

At training time the output sequence is shifted by one token and is used as input to predict all next tokens of the output sequence at once. At inference time, the model uses maximum likelihood estimation to select the next token from the final distribution D' (greedy decoding), as shown in Figure 2.

We further increase capacity of our generative model by learning to combine several individual distributions (D'_1 and D'_2 in Figure 2). The final distribution D' is then produced as a weighted sum of the intermediate distributions: $D' = \sum_{i=0}^{m} \alpha_i D'_i$ ($m = 2$ in our experiments). To produce $D'_i \in \mathbb{R}^{|V|}$ we pass the last hidden state of the Transformer Decoder $h \in \mathbb{R}^d$ through a separate linear layer for each intermediary distribution: $D'_i = W_i^H h + b^H$, where W_i^H is the weight matrix and b^H is the bias. For the weighting coefficients α_i we use the matrix of input embeddings $X \in \mathbb{R}^{n \times d}$, where n is the maximum sequence length and d is the embedding dimension, and the output of the first attention head of the Transformer Decoder $G \in \mathbb{R}^{n \times d}$ put through a layer normalization function: $\alpha_i = W_i^G norm(G) + W_i^X X + b_i^\alpha$, where all W are the weight matrices and b_i^α is the bias.

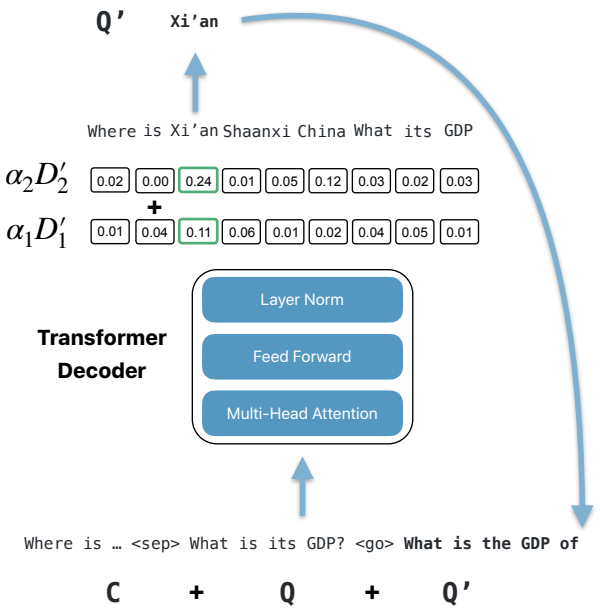


Figure 2: The question rewriting component uses the Transformer Decoder architecture, to recursively generate the tokens of an "explicit" question. At inference time, the generated output is appended to the input sequence for the next timestep in the sequence.

4.2 Retrieval QA Model

In the retrieval QA settings, the task is to produce a ranked list of text passages from a collection, ordered by their relevance to a given a natural language question [6, 22]. We employ the state-of-the-art approach to retrieval QA, which consists of two phases: candidate

selection and passage re-ranking. This architecture holds the state-of-the-art for the passage retrieval task on the MS MARCO dataset according to the recent experiments conducted by Xiong et al. [39].

In the first phase, a traditional retrieval algorithm (BM25) is used to quickly sift through the indexed collection retrieving top-k passages ranked by relevance to the input question Q'. In the second phase, a more computationally-expensive model is used to re-rank all question-answer candidate pairs formed using the previously retrieved set of k passages.

For re-ranking, we use a binary classification model that predicts whether the passage answers a question, i.e., the output of the model is the relevance score in the interval $[0, 1]$. The input to the re-ranking model is the concatenated question and passage with a separation token in between (see Figure 3 for the model overview). The model is initialized with weights learned from unsupervised pre-training on the language modeling (masked token prediction) task (BERT) [5]. During fine-tuning, the training objective is to reduce cross-entropy loss, using relevant passages and non-relevant passages from the top-k candidate passages.

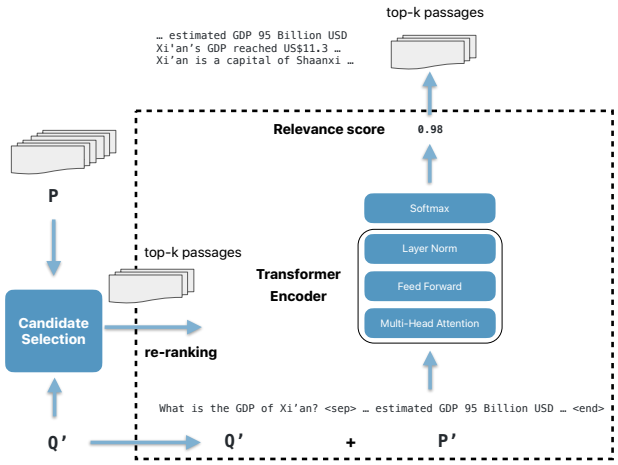


Figure 3: Retrieval QA component includes two sequential phases: candidate selection (BM25) followed by passage re-ranking (Transformer Encoder).

4.3 Extractive QA Model

The task of extractive QA is given a natural language question and a single passage find an answer as a contiguous text span within the given passage [29]. Our model for extractive QA consists of a Transformer-based bidirectional encoder (BERT) [5] and an output layer predicting the answer span. This type of model architecture corresponds to the current state-of-the-art setup for several reading comprehension benchmarks [15, 19].

The input to the model is the sequence of tokens formed by concatenating a question and a passage separated with a special $[SEP]$ token. The encoder layers are initialized with the weights of a Transformer model pre-trained on an unsupervised task (masked token prediction). The output of the Transformer encoder is a hidden vector T_i for each token i of the input sequence.

For fine-tuning the model on the extractive QA task, we add weight matrices W^s, W^e and biases b^s, b^e that produce two probability distributions over all the tokens of the given passage separately for the start (S) and end position (E) of the answer span. For each token i the output of the Transformer encoder T_i is passed through a linear layer, followed by a softmax normalizing the output logits over all the tokens into probabilities:

$$S_i = \frac{e^{W^s \cdot T_i + b^s}}{\sum_{j=1}^{n} e^{W^s \cdot T_j + b^s}} \qquad E_i = \frac{e^{W_e \cdot T_i + b_e}}{\sum_{j=1}^{n} e^{W^e \cdot T_j + b^e}} \tag{1}$$

The model is then trained to minimize cross-entropy between the predicted start/end positions (S_i and E_i) and the correct ones from the ground truth (y_S and y_E are one-hot vectors indicating the correct start and end tokens of the answer span):

$$loss = -\sum_{i=1}^{n} y_S \log S_i - \sum_{i=1}^{n} y_E \log E_i \tag{2}$$

At inference time all possible answer spans from position i to position j, where $j \geq i$, are scored by the sum of end and start positions' probabilities: $S_i + E_j$. The output of the model is the maximum scoring span (see Figure 4 for the model overview).

21% of the CANARD (QuAC) examples are Not Answerable (NA) by the provided passage. To enable our model to make No Answer predictions we prepend a special $[CLS]$ token to the beginning of the input sequence. For all No Answer samples we set both the ground truth start and end span positions to this token's position (0). Likewise, at inference time, predicting this special token is equivalent to a No Answer prediction for the given example.

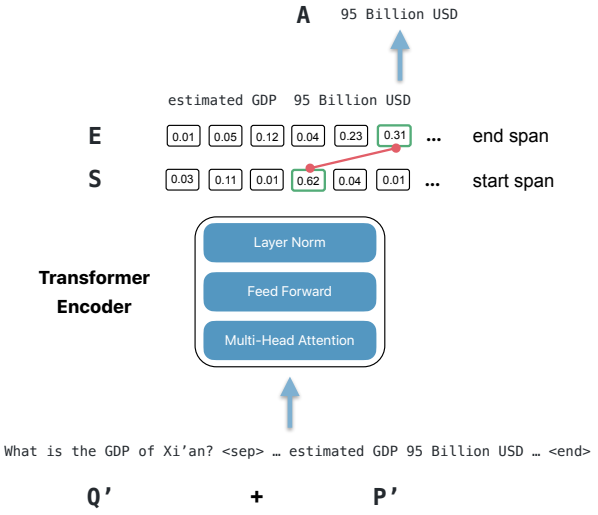


Figure 4: Extractive QA component predicts a span of text in the paragraph P', given an input sequence with the question Q' and passage P'.

5 EXPERIMENTAL SETUP

To evaluate the question rewriting approach we perform a range of experiments with several existing state-of-the-art QA models. In

Table 1: Datasets used for training and evaluation (with the number of questions).

Tasks	Question Rewriting	Retrieval QA	Extractive QA
Train	CANARD (35k)	MS MARCO (399k)	+MultiQA (75k) CANARD (35k)
Test	TREC CAsT (173) CANARD (5.5k)	TREC CAsT (173)	CANARD (5.5k)

the following subsections we describe the datasets used for training and evaluation, the set of metrics for each of the components, our baselines and details of the implementation. Our experimental setup is designed to evaluate gains in performance, reuse and traceability from introducing the QR component:

RQ1*: How does the proposed approach perform against competitive systems (performance)?*

RQ2*: How does non-conversational pre-training benefit the models with and without QR (reuse)?*

RQ3*: What is the proportion of errors contributed by each of the components (traceability)?*

5.1 Datasets

We chose two conversational QA datasets for the evaluation of our approach: (1) CANARD, derived from Question Answering in Context (QuAC) for extractive conversational QA [2], and (2) TREC CAsT for retrieval conversational QA [4]. See Table 1 for the overview of the datasets. Since TREC CAsT is relatively small we used only CANARD for training QR. The same QR model trained on CANARD is evaluated on both CANARD and TREC CAsT.

CANARD [7] contains sequences of questions with answers as text spans in a given Wikipedia article. It was built upon the QuAC dataset [2] by employing human annotators to rewrite original questions from QuAC dialogues into explicit questions. CANARD consists of 40.5k pairs of question rewrites that can be matched to the original answers in QuAC. We use CANARD splits for training and evaluation. We use the question rewrites provided in CANARD and articles with correct answer spans from QuAC. In our experiments, we refer to this joint dataset as CANARD for brevity.

To further boost performance of the extractive QA model we reuse the approach from Fisch et al. [8] and pre-train the model on MultiQA dataset, which contains $75k$ QA pairs from six standard QA benchmarks, and then fine-tune it on CANARD to adapt for the domain shift. Note that MultiQA is a non-conversational QA dataset.

TREC CAsT contains sequences of questions with answers to be retrieved as text passages from a large collection of passages: MS MARCO [22] with 8.6M passages + TREC CAR [6] with 29.8M passages. The official test set that we used for evaluation contains relevance judgements for 173 questions across 20 dialogues (topics). We use the model from Nogueira and Cho [23] for retrieval QA, which was tuned on a sample from MS MARCO with 12.8M query-passage pairs and 399k unique queries.

5.2 Metrics

Our evaluation setup corresponds to the one used in TREC CAsT, which makes our experiments directly comparable to the official TREC CAsT results. Mean average precision (*MAP*), mean reciprocal rank (*MRR*), normalized discounted cumulative gain (*NDCG@3*) and precision on the top-passage (*P@1*) evaluate quality of passage retrieval. Top-1000 documents are considered per query with a relevance judgement value cut-off level of 2 (the range of relevance grades in TREC CAsT is from 0 to 4).

We use *F1* and Exact Match (*EM*) for extractive QA, which measure word token overlap between the predicted answer span and the ground truth. We also report accuracy for questions without answers in the given passage (*NA Acc*).

Our analysis showed that *ROUGE* recall calculated for unigrams (ROUGE-1 recall) correlates with the human judgement of the question rewriting performance (Pearson 0.69), which we adopt for our experiments as well. ROUGE [18] is a standard metric of lexical overlap, which is often used in text summarization and other text generation tasks. We also calculate question similarity scores with the Universal Sentence Encoder (*USE*) model [1] (Pearson 0.71).

5.3 QR Baselines

The baselines were designed to challenge the need for a separate QR component by incorporating previous turns as direct input to custom QA components. Manual rewrites by human annotators provide the upper-bound performance for a QR approach and allows for an ablation study of the down-stream QA components.

Original. Original questions from the conversational QA datasets without any question rewriting.

Original + k-DT.. Our baseline approach for extractive QA prepends the previous k questions to the original question to compensate for the missing context. The questions are separated with a special token and used as input to the Transformer model. We report the results for $k = \{1, 2, 3\}$.

Original + k-DT*. Since in the first candidate selection phase we use BM25 retrieval function which operates on a bag-of-words representation, we modify the baseline approach for retrieval QA as follows. We select keywords from k prior conversation turns (not including current turn) based on their inverse document frequency (IDF) scores and append them to the original question of the current turn. We use the keyword-augmented query as the search query for **Anserini** [40], a Lucene toolkit for replicable information retrieval research, and if we use BERT re-ranking we concatenate the keyword-augmented query with the passages retrieved from the keyword-augmented query. We use the keywords with IDF scores above the threshold of 0.0001, which was selected based on a 1 million document sample of the MS MARCO corpus.

Human. To provide an upper bound, we evaluate all our models on the question rewrites manually produced by human annotators.

5.4 QR Models

In addition to the baselines described above, we chose several alternative models for question rewriting of the conversational context: (1)

co-reference resolution as in the TREC CAsT challenge; (2) PointerGenerator used for question rewriting on CANARD by Elgohary et al. [7] but not previously evaluated on the end-to-end conversational QA task; (3) CopyTransformer extension of the PointerGenerator model that replaces the bi-LSTM encoder-decoder architecture with a Transformer Decoder model. All models, except for co-reference, were trained on the train split of the CANARD dataset. Question rewrites are generated turn by turn for each dialogue recursively using already generated rewrites as previous turns. This is the same setup as in the TREC CAsT evaluation.

Co-reference. Anaphoric expressions in original questions are replaced with their antecedents from the previous dialogue turns. Co-reference dependencies are detected using a publicly available neural co-reference resolution model that was trained on OntoNotes [16].[1]

PointerGenerator. A sequence-to-sequence model for text generation with bi-LSTM encoder and a pointer-generator decoder [7].

CopyTransformer. The Transformer decoder, which, similar to pointer-generator model, uses one of the attention heads as a pointer [10]. The model is initialized with the weights of a pre-trained GPT2 model [28, 38] (Medium-sized GPT-2 English model: 24-layer, 1024-hidden, 16-heads, 345M parameters) and then fine-tuned on the question rewriting task.

Transformer++. The Transformer-based model described in Section 4.1. Transformer++ is initialized with the weights of the pre-trained GPT2 model, same as in CopyTransformer.

5.5 QA Models

Our retrieval QA approach is implemented as proposed in [23] using Anserini for the candidate selection phase with BM25 (top-1000 passages) and $BERT_{LARGE}$ for the passage re-ranking phase (Anserini + BERT). Both components were fine-tuned only on the MS MARCO dataset ($k_1 = 0.82, b = 0.68$).[2]

We train several models for extractive QA on different variants of the training set based on the CANARD training set [7]. All models are first initialized with the weights of the $BERT_{LARGE}$ model pre-trained using the whole word masking [5].

CANARD-O.. The baseline models were trained using original (implicit) questions of the CANARD training set with a dialogue context of varying length (Original and Original + k-DT). The models are trained separately for each $k = \{0, 1, 2, 3\}$, where $k = 0$ corresponds to the model trained only on the original questions without any previous dialogue turns.

CANARD-H.. To accommodate input of the question rewriting models, we train a QA model that takes human rewritten question from the CANARD dataset as input without any additional conversation context, i.e., as in the standard QA task.

MultiQA → CANARD-H.. Since the setup with rewritten questions does not differ from the standard QA task, we experiment with pretraining the extractive QA model on the MultiQA dataset with explicit questions [8], using parameter choices introduced by Longpre et al. [20]. We further fine-tune this model on the target CANARD dataset to better adopt it for the domain shift in CANARD QA samples (see Figure 5).

[1] https://github.com/kentonl/e2e-coref
[2] https://github.com/nyu-dl/dl4marco-bert

Table 2: Evaluation results of the QR models. *Human performance is measured as the difference between two independent annotators' rewritten questions, averaged over 100 examples. This provides an estimate of the upper bound.

Test Set	Question	ROUGE	USE	EM
CANARD	Original	0.51	0.73	0.12
	Co-reference	0.68	0.83	0.48
	PointerGenerator	0.75	0.83	0.22
	CopyTransformer	0.78	0.87	0.56
	Transformer++	**0.81**	**0.89**	**0.63**
	Human*	0.84	0.90	0.33
TREC CAsT	Original	0.67	0.80	0.28
	Co-reference	0.71	0.80	0.13
	PointerGenerator	0.71	0.82	0.17
	CopyTransformer	0.82	0.90	0.49
	Transformer++	**0.90**	**0.94**	**0.58**
	Human*	1.00	1.00	1.00

6 RESULTS

RQ1*: How does the proposed approach perform against competitive systems (performance)?* Our approach, using question rewriting for conversational QA, consistently outperforms the baselines that use previous dialogue turns, in both retrieval and extractive QA tasks (see Tables 3-5). Moreover, it shows considerable improvement over the latest results reported on the TREC CAsT dataset (see Table 4).

The precision-recall trade-off curve in Figure 6 shows that question rewriting performance is close to the performance achieved by manually rewriting implicit questions. Our results also indicate that the QR performance metric is able to correctly predict the model that performs consistently better across both QA tasks (see Table 2).

Passage re-ranking with BERT always improves ranking results (almost a two-fold increase in MAP, see Table 3). Keyword-based baselines (Original + k-DT*) prove to be very strong outperforming both Co-reference and PointerGenerator models on all three performance metrics. Both MRR and NDCG@3 are increasing with the number of turns used for sampling keywords, while MAP is slightly decreasing, which indicates that it brings more relevant results at the very top of the rank but non-relevant results also receive higher scores. In contrast, the baseline results for Anserini + BERT model indicate that the re-ranking performance for all metrics decreases if the keywords from more than 2 previous turns are added to the original question.

Similarly for extractive QA, the model incorporating previous turns proved to be a very strong baseline (see Table 5). The performance results also suggest that all models do not discriminate well the passages that do not have an answer to the question (71% accuracy on the human rewrites). We notice that the baseline models, which were trained with the previous conversation turns, tend to adopt a conservative strategy by answering more questions as

Table 3: Retrieval QA results on the TREC CAsT test set.

QA Input	QA Model	MAP	MRR	NDCG@3
Original	Anserini	0.089	0.245	0.131
Original + 1-DT*		0.133	0.343	0.199
Original + 2-DT*		0.130	0.374	0.213
Original + 3-DT*		0.127	0.396	0.223
Co-reference		0.109	0.298	0.172
PointerGenerator		0.100	0.273	0.159
CopyTransformer		0.148	0.375	0.213
Transformer++		0.190	0.441	0.265
Human		0.218	0.500	0.315
Original	Anserini	0.172	0.403	0.265
Original + 1-DT*	+BERT	0.230	0.535	0.378
Original + 2-DT*		0.245	0.576	0.404
Original + 3-DT*		0.238	0.575	0.401
Co-reference		0.201	0.473	0.316
PointerGenerator		0.183	0.451	0.298
CopyTransformer		0.284	0.628	0.440
Transformer++		**0.341**	**0.716**	**0.529**
Human		0.405	0.879	0.589

Table 4: Comparison with the state-of-the-art results reported on the TREC CAsT test set.

Approach	NDCG@3
Mele et al. [21]	0.397
Voskarides et al. [37]	0.476
Yu et al. [41]	0.492
Ours	**0.529**

"Unanswerable" (NA). Controlling for this effect, we show that our QR model gains a higher performance level by actually answering questions that have answers.

RQ2*: How does non-conversational pre-training benefit the models with and without QR (reuse)?* We observe that pre-training on MultiQA improves performance of all extractive QA model. However, it is much more prominent for the systems using question rewrites (7% increase in EM and 6% in F1 when using human-rewritten questions). The models do not require any additional fine-tuning, when using QR, since the type of input to the QA model remains the same (non-conversational). While for Original + k-DT models fine-tuning is required also to adopt the model for the new type of input data (conversational). Note that, in this case, we had to fine-tune all models but for another reason. The style of questions in CANARD is rather different from other QA datasets in MultiQA. Figure 5 demonstrates that a small portion of the training data is sufficient to adopt a QR-based model trained with non-conversational samples to work well on CANARD.

RQ3*: What is the proportion of errors contributed by each of the components (traceability)?* We measure the effect from question rewriting for each of the questions by comparing the answers produced for the original, the model-rewritten (Transformer++) and the human-rewritten question (see Tables 6-7). This approach allows us to pinpoint the cases, in which QR contributes to the QA performance, and distinguish them from cases in which the answer can be found using the original question as well.

Assuming that humans always produce correct question rewrites, we can attribute all cases in which these rewrites did not result in a correct answer as errors of the QA component (rows 1-4 in Tables 6-7). The next two rows 5-6 show the cases, where human rewrites succeeded but the model rewrites failed, which we consider to be a likely error of the QR component. The last two rows are true positives for our model, where the last row combines cases where the original question was just copied without rewriting (numbers in brackets) and other cases when rewriting was not required. Since

Table 5: Extractive QA results on the CANARD test set. F1 and EM is calculated for both answerable and unanswerable questions, while NA Acc only for unanswerable questions.

QA Input	Training Set	EM	F1	NA Acc
Original	CANARD-O	38.68	53.65	66.55
Original + 1-DT		42.04	56.40	66.72
Original + 2-DT		41.29	56.68	68.11
Original + 3-DT		42.16	56.20	68.72
Original	CANARD-H	39.44	54.02	65.42
Human		42.36	57.12	68.20
Original	MultiQA →	41.32	54.97	65.84
Original + 1-DT	CANARD-H	43.15	57.03	68.64
Original + 2-DT		42.20	57.33	69.42
Original + 3-DT		43.29	57.87	**71.50**
Co-reference		42.70	57.59	66.20
PointerGenerator		41.93	57.37	63.16
CopyTransformer		42.67	57.62	68.02
Transformer++		**43.39**	**58.16**	68.29
Human		45.40	60.48	70.55

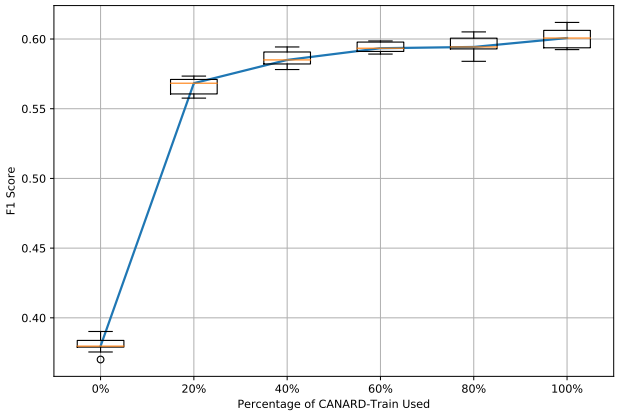


Figure 5: Effect from fine-tuning the MultiQA model on a portion of the target CANARD-H dataset due to the domain shift between the datasets.

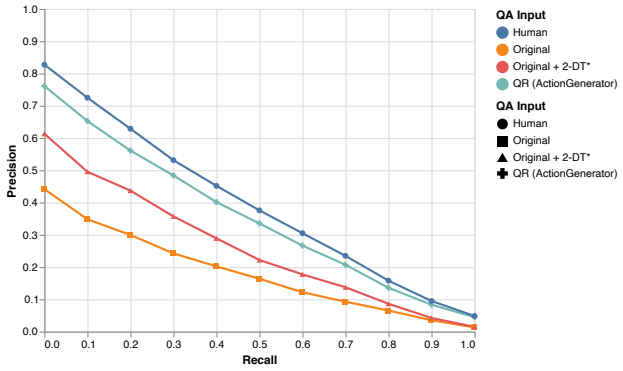


Figure 6: Precision-recall curve illustrating model performance on the TREC CAsT test set for Anserini + BERT.

there is no single binary measure for the answer correctness, we select different cut-off thresholds for our QA metrics.

The majority of errors stem from the QA model: 29% of the test samples for retrieval and 55% for extractive estimated for P@1 and F1, comparing to 11% and 5% for QR respectively. Note that it is a rough estimate since we cannot automatically distinguish the cases that failed both QA and QR.

Overall, we observe that the majority of questions in extractive QA setup can be correctly answered without rewriting or accessing the conversation history. In other words, the extractive QA model tends to return an answer even when given an incomplete ambiguous question. This finding also explains the low NA Acc results reported in Table 5. Our results provide evidence of the deficiency of the reading comprehension setup, which was also reported in the previous studies on non-conversational datasets [13, 17, 32].

In contrast, in the retrieval QA setup, only 10% of the questions in TREC CAsT were rewriten by human annotators that did not need rewriting to retrieve the correct answer. These results highlight the difference between the retrieval and extractive QA training/evaluation setup. More details on our error analysis approach and results can be found in Vakulenko et al. [36].

7 CONCLUSION

We showed in an end-to-end evaluation that question rewriting is effective in extending standard QA approaches to a conversational setting. Our results set the new state-of-the-art on the TREC CAsT 2019 dataset. The same QR model also shows superior performance on the answer span extraction task evaluated on the CANARD/QuAC dataset. Based on the results of our analysis, we conclude that QR is a challenging but also very promising task that can be effectively implemented into conversational QA approaches.

We also confirmed that the QR metric provides a good indicator for the end-to-end QA performance. Thereby, it is reliable to be used for the QR model selection, which would help to avoid more costly end-to-end evaluation in the future.

The experimental evaluation and the detailed analysis we provide increases our understanding of the main error sources. QR performance is sufficiently high on both CANARD and TREC CAsT datasets, while the QA performance even given human rewritten

Table 6: Break-down analysis of all retrieval QA results for the TREC CAsT dataset. Each row represents a group of QA samples that exhibit similar behaviour. ✓ indicates that the answer produced by the QA model was correct or × – incorrect, according to the thresholds provided in the right columns. We consider three types of input for every QA sample: the question from the test set (Original), generated by the best QR model (Transformer++) or rewritten manually (Human). The numbers correspond to the count of QA samples for each of the groups. The numbers in parenthesis indicate how many questions do not require rewriting, i.e., should be copied from the original.

			P@1	NDCG@3		
Original	QR	Human	= 1	> 0	≥ 0.5	= 1
×	×	×	49 (14)	10 (1)	55 (20)	154 (49)
✓	×	×	0	0	0	0
×	✓	×	2	0	1	0
✓	✓	×	0	1	1	0
×	×	✓	19	10	25	4
✓	×	✓	0	1	0	0
×	✓	✓	48	63	47	11
✓	✓	✓	55 (37)	88 (52)	44 (33)	4 (4)
			Total			173 (53)

Table 7: Break-down analysis of all extractive QA results for the CANARD dataset, similar to Table 6.

Original	QR	Human	F1 > 0	F1 ≥ 0.5	F1 = 1
×	×	×	847 (136)	1855 (235)	2701 (332)
✓	×	×	174	193	181
×	✓	×	19	35 (2)	40 (1)
✓	✓	×	135	153	120
×	×	✓	141	288	232
✓	×	✓	65 (1)	57 (1)	40
×	✓	✓	226	324	269
✓	✓	✓	3964 (529)	2666 (428)	1988 (333)
			Total		5571 (666)

questions for both tasks lags behind. This result suggest that the major improvement in conversational QA will come from improving the standard QA models.

In future work we would like to evaluate QR performance with the joint model integrating passage retrieval and answer span extraction. Recent results reported by Qu et al. [26] indicate that there is sufficient room for improvement in the history modeling phase.

On the other hand, the QR-QA type of architecture is generic enough to incorporate other types of context, such as a user model or an environmental context obtained from multi-modal data (deictic reference). Experimental evaluation of QR-QA performance augmented with such auxiliary inputs is a promising direction for future work.

ACKNOWLEDGEMENTS

We would like to thank our colleagues Srinivas Chappidi, Bjorn Hoffmeister, Stephan Peitz, Russ Webb, Drew Frank, and Chris DuBois for their insightful comments.

REFERENCES

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 169–174.

[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2174–2184.

[3] Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 729–738.

[4] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. CAsT 2019: The Conversational Assistance Track Overview. In *Proceedings of the 28th Text REtrieval Conference*. 13–15.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[6] Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. 2018. TREC Complex Answer Retrieval Overview. TREC.

[7] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 5920–5926.

[8] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. *arXiv preprint arXiv:1910.09753* (2019).

[9] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[10] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4098–4109.

[11] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. 2015. Context- and Content-aware Embeddings for Query Rewriting in Sponsored Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*. 383–392.

[12] Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2018. Dialog-to-Action: Conversational Question Answering Over a Large-Scale Knowledge Base. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*. 2946–2955.

[13] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2021–2031.

[14] Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on Conversational Question Answering. *CoRR* abs/1909.10772 (2019).

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[16] Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 687–692.

[17] Mike Lewis and Angela Fan. 2019. Generative Question Answering: Learning to Answer the Whole Question. In *7th International Conference on Learning Representations*.

[18] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 220–227.

[21] Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2020. Topic Propagation in Conversational Search. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 2057–2060.

[22] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

[23] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[24] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019).

[25] Yannis Papakonstantinou and Vasilis Vassalos. 1999. Query Rewriting for Semistructured Data. In *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data*. 455–466.

[26] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-Retrieval Conversational Question Answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 539–548.

[27] Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019. Attentive History Selection for Conversational Question Answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1391–1400.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019).

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[30] Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling Multi-Domain Dialogue State Tracking via Query Reformulation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 97–105.

[31] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.

[32] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 856–865.

[33] Shariq Rizvi, Alberto O. Mendelzon, S. Sudarshan, and Prasan Roy. 2004. Extending Query Rewriting Techniques for Fine-Grained Access Control. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France, June 13-18, 2004*. 551–562.

[34] Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. *arXiv preprint arXiv:1906.07004* (2019).

[35] Andrew L Thomas. 1979. Ellipsis: The Interplay of Sentence Structure and Context. *Lingua Amsterdam* 47, 1 (1979), 43–68.

[36] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering. *Proceedings of the 2020 EMNLP Workshop SCAI: The 5th International Workshop on Search-Oriented Conversational AI* (2020).

[37] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query Resolution for Conversational Search with Limited Supervision. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 921–930.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771* (2019).

[39] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[40] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th International Conference on Research and Development in Information Retrieval*. 1253–1256.

[41] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1933–1936.

2. Open-Domain Question Answering Goes Conversational via Question Rewriting

Bibliographic Information

Raviteja Anantha*, **Svitlana Vakulenko***, Zhucheng Tu, Shayne Longpre, Stephen Pulman, Srinivas Chappidi: Open-Domain Question Answering Goes Conversational via Question Rewriting. NAACL-HLT 2021: 520-534.

Open-Domain Question Answering Goes Conversational via Question Rewriting

Raviteja Anantha[★♦], **Svitlana Vakulenko**[★♠♥], **Zhucheng Tu**[♦], **Shayne Longpre**[♦], **Stephen Pulman**[♦], **Srinivas Chappidi**[♦]

[♦]Apple Inc.

[♠]University of Amsterdam, Amsterdam, the Netherlands

Abstract

We introduce a new dataset for **Q**uestion **Re**writing in **C**onversational **C**ontext (QReCC), which contains 14K conversations with 80K question-answer pairs. The task in QReCC is to find answers to conversational questions within a collection of 10M web pages (split into 54M passages). Answers to questions in the same conversation may be distributed across several web pages. QReCC provides annotations that allow us to train and evaluate individual subtasks of question rewriting, passage retrieval and reading comprehension required for the end-to-end conversational question answering (QA) task. We report the effectiveness of a strong baseline approach that combines the state-of-the-art model for question rewriting, and competitive models for open-domain QA. Our results set the first baseline for the QReCC dataset with F1 of 19.10, compared to the human upper bound of 75.45, indicating the difficulty of the setup and a large room for improvement.

1 Introduction

It is often not possible to address a complex information need with a single question. Consequently, there is a clear need to extend open-domain question answering (QA) to a conversational setting. This task is commonly referred to as conversational (interactive or sequential) QA (Webb, 2006; Saeidi et al., 2018; Reddy et al., 2019). Conversational QA requests an answer conditioned on both the question and the previous conversation turns as context. Previously proposed large-scale benchmarks for conversational QA, such as QuAC and CoQA, limit the topic of conversation to the content of a single document. In practice, however, the answers can be distributed across several documents that are relevant to the conversation, or the topic of the conversation may also drift. To investigate this phenomena and develop approaches suitable for the complexities of this task, we introduce a new dataset for open-domain conversational QA, called QReCC.[1] The dataset consists of 13.6K conversations with an average of 6 turns per conversation.

Question: Tell me about the benefits of **Yoga**?
Answer: Increased flexibility, muscle strength…
URL: https://osteopathic.org/what-is-osteopathic-medicine/benefits-of-yoga

Question: Does **it** help in reducing stress?
Rewrite: Does **Yoga** help in reducing stress?
Answer: Yoga may help reduce stress, lower blood pressure, and lower your heart rate.
URL: https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/yoga/art-20044733

Question: What are some of the main types?
Rewrite: What are some of the main types **of Yoga**?
Answer: Hatha, Kundalini, Ashtanga, …
URL: https://www.mindbodygreen.com/articles/the-11-major-types-of-yoga-explained-simply

Question: What are common poses in Kundalini Yoga?
Rewrite: What are common poses in Kundalini Yoga?
Answer: Lotus Pose, Celibate Pose, Perfect Pose, …
URL: https://www.kundaliniyoga.org/Asanas

Figure 1: A snippet of a sample conversation from QReCC with question rewrites and answer provenance links. **Orange** indicates coreference cases where the highlighted token should be replaced with its antecedent (in **bold**). **Blue** indicates the tokens that should be generated to make the question unambiguous outside of the conversational context.

A conversation in QReCC consists of a sequence of question-answer pairs. The answers to questions were produced by human annotators, who looked up relevant information on the web using a search engine. QReCC is therefore the first large-scale dataset for conversational QA that incorporates an information retrieval subtask. QReCC is accompanied with scripts for building a collection of passages from the Common Crawl and the Wayback Machine for passage retrieval.

QReCC is inspired by the task of question rewriting (QR) that allows us to reduce the task of conversational QA to non-conversational QA by

★ Equal contribution.
♥ Work done as an intern at Apple Inc.

[1] https://github.com/apple/ml-qrecc

generating self-contained versions of contextually-dependent questions. QR was recently shown crucial for porting retrieval QA architectures to a conversational setting (Dalton et al., 2019). Follow-up questions in conversational QA often depend on the previous conversation turns due to ellipsis (missing content) and coreference (anaphora). Every question-answer pair in QReCC is also annotated with a question rewrite. We evaluate the quality of these rewrites as self-contained questions in terms of the ability of the rewritten question, when used as input to the web search engine, to retrieve the correct answer. A snippet of a sample QReCC conversation is given in Figure 1.

The dataset collection included two phases: (1) dialogue collection, and (2) document collection. First, we set up an annotation task to collect dialogues with question-answer pairs along with question rewrites and answer provenance links. Second, after all dialogues were collected we downloaded the web pages using the provenance links, and then extended this set with a random sample of other web pages from Common Crawl, preprocessed and split the pages into passages.

To produce the first baseline, we augment an open-domain QA model with a QR component that allows us to extend it to a conversational scenario. We evaluate this approach on the QReCC dataset, reporting the end-to-end effectiveness as well as the effectiveness on the individual subtasks separately.

Our contributions. We collected the first large-scale dataset for end-to-end, open-domain conversational QA that contains question rewrites that incorporate conversational context. We present a systematic comparison of existing automatic evaluation metrics on assessing the quality of question rewrites and show the metrics that best correlate with human judgement. We show empirically that QR provides a unified and effective solution for resolving references — both co-reference and ellipsis — in multi-turn dialogue setting and positively impacts the conversational QA task. We evaluate the dataset using a baseline that incorporates the state-of-the-art model in QR and competitive models for passage retrieval and answer extraction. This dataset provides a resource for the community to develop, evaluate, and advance methods for end-to-end, open-domain conversational QA.

2 Related Work

QReCC builds upon three publicly available datasets and further extends them to the open-domain conversational QA setting: Question Answering in Context (QuAC) (Choi et al., 2018), TREC Conversational Assistant Track (CAsT) (Dalton et al., 2019) and Natural Questions (NQ) (Kwiatkowski et al., 2019). QReCC is the first large-scale dataset that supports the tasks of QR, passage retrieval, and reading comprehension (see Table 1 for the dataset comparison).

Open-domain QA. Reading comprehension (RC) approaches were recently extended to incorporate a retrieval subtask (Chen et al., 2017; Yang et al., 2019; Lee et al., 2019). This task is also referred to as machine reading at scale (Chen et al., 2017) or end-to-end QA (Yang et al., 2019). In this setup a reading comprehension component is preceded by a document retrieval component. The answer spans are extracted from documents retrieved from a document collection, given as input. The standard approach to end-to-end open-domain QA is (1) use an efficient filtering approach to reduce the number of candidate passages to the top-k of the most relevant ones (usually BM25 based on the bag-of-words representation); and then (2) re-rank the subset of the top-k relevant passages using a more fine-grained approach, such as BERT based on vector representations (Yang et al., 2019).

Conversational QA. Independently from end-to-end QA, the RC task was extended to a conversational setting, in which answer extraction is conditioned not only on the question but also on the previous conversation turns (Choi et al., 2018; Reddy et al., 2019). The first attempt at extending the task of information retrieval (IR) to a conversational setting was the recent TREC CAsT 2019 task (Dalton et al., 2019). The challenge was to rank passages from a passage collection by their relevance to an input question in the context of a conversation history. The size of the collection in CAsT 2019 was 38.4M passages, requiring efficient IR approaches. As efficient retrieval approaches operate on bag-of-words representations they need a different way to handle conversational context since they can not be trained end-to-end using a latent representation of the conversational context. A solution to this computational bottleneck was a QR model that learns to sample tokens from the conversational context as a pre-processing step before QA.

Table 1: The datasets that QReCC extends to open-domain conversational QA (QuAC, CAsT and NQ) and the datasets that are complementary to QReCC (CANARD and SaaC). RC - Reading Comprehension, PR - Passage Retrieval, QR - Question Rewriting.

Dataset	#Dialogues	#Questions	Task	Provenance
QuAC (Choi et al., 2018)	13.6K	98K	RC	-
NQ (Kwiatkowski et al., 2019)	0	307K	RC	-
CAsT (Dalton et al., 2019)	80	748	PR	-
CANARD (Elgohary et al., 2019)	5.6K	41K	QR	QuAC
OR-QuAC (Qu et al., 2020)	5.6K	41K	PR+RC	QuAC+CANARD
SaaC (Ren et al., 2020)	80	748	QR+PR+RC	CAsT
QReCC (our work)	13.7K	81K	QR+PR+RC	QuAC+NQ+CAsT

Question Rewriting. CANARD (Elgohary et al., 2019) provides rewrites for the conversational questions from the QuAC dataset. QR effectively modifies all follow-up questions such that they can be correctly interpreted outside of the conversational context as well. This extension to the conversational QA task proved especially useful while allowing retrieval models to incorporate conversational context (Voskarides et al., 2020; Vakulenko et al., 2020; Lin et al., 2020).

More recently, Qu et al. introduced OR-QuAC dataset that was automatically constructed from QuAC and CANARD datasets. OR-QuAC uses the same rewrites and answers as the ones provided in QuAC and CANARD. In contrast to OR-QuAC, the answers in QReCC are not tied to a single Wikipedia page. The answers can be distributed across several web pages. QReCC's passage collection is also larger and more diverse: 11M passages from Wikipedia in OR-QuAC vs. 54M passages from CommonCrawl in QReCC. The answers in OR-QuAC are single spans, whereas QReCC answers were produced by human annotators instructed to imitate natural conversational answers and may include several spans from different parts of the same web page.

TREC CAsT 2019 paved the way to conversational QA for retrieval but had several important limitations: (1) no training data and (2) no answer spans. First, the size of the CAsT dataset is limited to 80 dialogues, which is nowhere enough for training a machine-learning model. This was also the reason why CANARD played such an important role for the development of retrieval-based approaches even though it was collected as a RC dataset. Second, the task in TREC CAsT 2019 was conversational passage retrieval not extractive QA since the expected output was ranked passages and not a text span. We designed QReCC to overcome both of these limitations.

The size of the QReCC dataset is comparable with other large-scale conversational QA datasets (see Table 1). The most relevant to our work is the concurrent work by Ren et al., who extended the TREC CAsT dataset with crowd-sourced answer spans. Since the size of this dataset is inadequate for training a machine-learning model and can be used only for evaluation, the authors train their models on the MS MARCO dataset instead, which is a non-conversational QA dataset (Bajaj et al., 2016). Their evaluation results show how the performance degrades due to the lack of conversational training data. TREC CAsT will continue in the future and the QReCC dataset provides a valuable benchmark helping to train and evaluate novel conversational QA approaches.

3 Dialogue Collection

To simplify the data collection task we decided to use questions from pre-existing QA datasets as seeds for dialogues in QReCC. We used questions from QuAC, CAsT and NQ. While QuAC and CAsT datasets contain question sequences, NQ is not a conversational dataset but contains standalone questions from web search. We use the NQ dataset to increase and diversify the number of samples beyond QuAC and CAsT by generating more rewrites for cases beyond coreference resolution. The majority of the follow-up questions in QuAC require coreference resolution for QR. Therefore, we explicitly instructed the annotators to use NQ as a start of a conversation and then come up with relevant follow-up questions, which would require generation of missing content, i.e., ellipsis, instead of coreference resolution for QR.

The task for the annotators was also to answer questions using a web search engine. Question rewrites were used as input to a search engine. This

Table 2: Summary statistics for the QReCC dataset.

QReCC	Train	Dev.	Test	All
# questions (Qs)	50.8K	12.7K	16.4K	80.0K
# dialogues	8.7K	2.2K	2.8K	13.6K
max Qs/dialogue	12	12	12	12
avg Qs/dialogue	6	6	6	6
min Qs/dialogue	5	5	5	5
% replacement	53	52	53	52
% insertion	35	36	37	38
% copy	11	11	9	9
% removal	1	1	1	1

setup helps to obtain feedback on the quality of QR with respect to the effectiveness of answer retrieval (see Section 6 for more details on using search results for the evaluating QR). Finally, the question-answer pair is annotated with the link to the web page that was used to produce the answer.

Thereby, every dialogue was produced by the same annotator including the questions, answers and rewrites. This design decision is called *self-dialog technique* that was shown to help improve quality of the data by avoiding some of the challenges observed in simulated dialogues produced by pairs of annotators (Byrne et al., 2019).

A team of 30 professional annotators with a project lead were employed to perform the task. The annotation task was described in the guidelines (see Appendix B for more details). To ensure the quality of the annotations we followed a post-hoc evaluation procedure, in which 5 reviewers go through the dataset and update incorrect examples they identify with consensus.

4 Dialogue Analysis

QReCC contains 13,598 dialogues with 79,952 questions in total. 9.3K dialogues are based on the questions from QuAC; 80 are from TREC CAsT; and 4.4K are from NQ. 9% of questions in QReCC do not have answers. We still retained the question rewrites even if no answer was found on the web. 112 questions were annotated with links to web pages without answer texts, e.g. "May I have a link to road signs in Singapore?"

We prepared three standard dataset splits and ensured that they are balanced in terms of the standard dialogue statistics and the types of QR (see Table 2). We distinguish four types of QR. They differ with respect to the intervention required to resolve contextual dependencies in dialogue. These

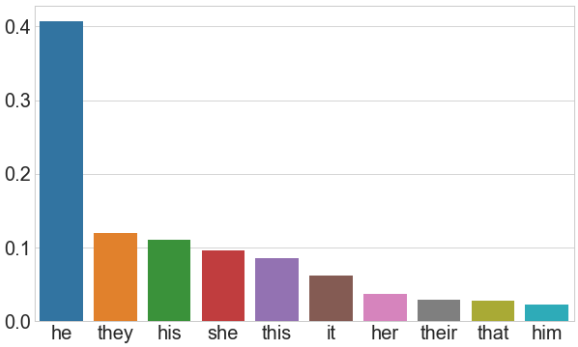


Figure 2: The 10 most frequently replaced tokens in QReCC.

types can be automatically identified by measuring the difference between an original question Q and a question rewrite R that are represented as sets using the bag-of-words:

- *Insertion* – new tokens are added to the original question to produce the rewrite (e.g., "What are some of the main types" → "What are some of the main types *of Yoga*?"):
  $Q \setminus R = \varnothing \wedge R \setminus Q \neq \varnothing$

- *Removal* – some tokens are removed from the question to produce the rewrite (e.g., "Can you tell me about the C++ language mentioned" → "Can you tell me about the C++ language"):
  $Q \setminus R \neq \varnothing \wedge R \setminus Q = \varnothing$

- *Replacement* – some tokens are added and some are removed to produce the rewrite (e.g., "Does it help in reducing stress" → "Does *Yoga* help in reducing stress"):
  $Q \setminus R \neq \varnothing \wedge R \setminus Q \neq \varnothing$

- *Copy* – no modification is needed, i.e., the original question is already contextually independent (e.g., "What are common poses in Kundalini Yoga?"):
  $Q \setminus R = \varnothing \wedge R \setminus Q = \varnothing$, i.e., $Q = R$

The majority of questions in QReCC (52%) require *Replacement*. Figure 2 shows the tokens that are most frequently replaced in QR. All of them are pronouns that require anaphora resolution. By specifically targeting more rare types of question rewriting in our data collection task we managed to increase the proportion of the *Insertion* cases in our dataset. This allows us to train and evaluate the ability of the model to reconstruct missing context, which cannot be achieved using traditional co-reference resolution approaches.

5 Document Collection

We download the web pages using the answer provenance links provided by the annotators from the Internet Archive Wayback Machine.[2] Then, we complement the relevant pages with randomly sampled web pages that constitute 1% of the Common Crawl dataset identified as English pages. The final collection consists of approximately 14K pages from the Wayback Machine and 9.9M random web pages from the Common Crawl dataset. The scripts for reproducing the passage collection are on GitHub. See Appendix A.2 for more details.

After downloading the pages we extract the textual content from the HTML and split texts into passages of least 220 tokens. After segmentation, we have a total of 54M passages which we index using Anserini (Yang et al., 2017).

We search the passage collection using the human annotated answers to augment the dataset with alternative sources of correct answers. For each document returned, we identify the span in the document that has the highest token overlap (F1) with the human answer. We consider all documents with F1 $\geq$ 0.8 as relevant. Verifying adequacy of this simple heuristic by human annotators is left for future work.

6 Question Rewriting Metrics Validation

BLEU has typically been used in previous work for measuring the quality of QR (Elgohary et al., 2019; Lin et al., 2020). We conduct a systematic evaluation and compare BLEU with alternative metrics, previously applied in summarization and translation, to ensure the most reliable metrics we can obtain for the model selection. Our evaluation shows that BLEU does not compare favourably with other metrics in evaluating the quality of QR.

Task. We took a random sample of 10K questions and used a seq-to-seq model (Nallapati et al., 2016) trained with questions and conversation context from the QReCC dataset to generate question rewrites. These generated rewrites were compared to the ground truth rewrites produced by human annotators. Different annotators graded each model-generated rewrite with a binary label: 0 (incorrect rewrite) or 1 (correct rewrite). For a question rewrite to be correct it does not have to exactly match the ground truth rewrite, but it should correctly capture the conversational context and be a self-contained question. For example, the model-generated rewrite "What are the global warming dangers?" is a correct rewrite with the ground truth rewrite being "What are the dangers of global warming?". In addition, we also assess the variance of the human assessments. The Pearson correlation between any two annotators on average is 0.94. We observed the mean and the variance to be 0.083 and 0.076 respectively. Performing a two-tail statistical significance test shows the P-value to be 0.0201.

We use several automated metrics to compare the rewrites with the ground truth and compute their Pearson correlation with the human judgements (see Table 3 for results).

Exact Match is a binary variable that indicates the token set overlap applied after the standard preprocessing: lower-casing, stemming, punctuation and stopword removal.

ROUGE (Lin, 2004) reflects similarity between two texts in terms of n-gram overlap (R-1 for unigrams; R-2 for bigrams and R-L for the longest common n-gram). We report the mean for precision (P), recall (R) and F-measure (F).

METEOR (Denkowski and Lavie, 2014) is a machine translation metric based on exact, stem, synonym, and paraphrase matches between words and phrases.

BLEU (Papineni et al., 2002) is a text similarity metric that uses a modified form of precision and n-grams from candidate and reference texts.

Embeddings group several unsupervised approaches that produce a sentence-level vector representation: Universal Sentence Encoder (Cer et al., 2018) and InferSent (Conneau et al., 2017).

Search Results – we use both question rewrites in Google Search and compare the overlap between the produced page ranks in terms of the standard IR metrics: Recall@k for the top-k links, Average Recall (AR) and Normalized Discounted Cumulative Gain (NDCG).

The best performing metric in our experiments (i.e., closest to the human judgement) is the set

[2] We use the version of a web page, which is the closest to the end date of the dialogue collection (November 24, 2019).

Table 3: Comparison of different evaluation metrics in terms of Pearson correlation with the human judgment of the question rewriting quality.

Metrics		Pearson	Metrics	Pearson
Exact Match		0.56	ROUGE-1 P	0.51
Embeddings	**USE**	**0.67**	**ROUGE-1 R**	**0.63**
	InferSent	0.48	ROUGE-1 F	0.61
Search Results	R@1	0.66	ROUGE-2 P	0.54
	R@2	0.72	ROUGE-2 R	0.57
	R@3	0.73	ROUGE-2 F	0.57
	R@4	0.74	ROUGE-L P	0.50
	R@5	0.77	ROUGE-L R	0.61
	R@10	**0.80**	ROUGE-L F	0.58
	AR	0.79	METEOR	0.59
	NDCG	0.74	BLEU	0.58

Table 4: Evaluation results of QR models (mean with 95% confidence intervals). *Human QR metrics are computed across 5 different random samples of 1000 question rewrites from the intersection of QReCC and CANARD conversations.

Model/Metrics	ROUGE-1 R	USE	R@10
AllenAI Coref (Lee et al., 2018)	67.1% ± 10E-4%	82.3% ± 10E-3%	56.1% ± 10E-4%
Generator (Radford et al., 2019)	73.4% ± 0.6%	86.2% ± 0.9%	69.1% ± 0.2%
Generator + Multiple-choice (Wolf et al., 2019b)	74.1% ± 0.5%	86.3% ± 0.4%	70.2% ± 0.1%
PointerGenerator (Elgohary et al., 2019)	80.2% ± 0.8%	89.1% ± 1.1%	75.3% ± 0.3%
GECOR (Quan et al., 2019)	84.1% ± 0.3%	91.8% ± 0.2%	78.1% ± 0.2%
CopyTransformer (Gehrmann et al., 2018)	86.1% ± 0.5%	92.8% ± 0.3%	79.4% ± 0.3%
Transformer++	**89.5% ± 0.4%**	**95.2% ± 0.2%**	**83.2% ± 0.3%**
Human*	94.6% ± 0.2%	97.3% ± 0.1%	87.2% ± 0.1%

overlap of the web search results (**R@10**). The best metrics independent of QA are Universal Sentence Embedding (**USE**) and unigram recall (**ROUGE-1 R**). We provide more details of the metrics performance illustrated with examples and the discussion in Appendix C. We use the set of all three best evaluation metrics to select the optimal QR model for our baseline approach.

7 Baseline Approach

We extend BERTserini (Yang et al., 2019), an efficient approach to open-domain QA, with a QR model to incorporate conversational context. This approach consists of three stages: (1) QR, (2) PR and (3) RC. First, a model is trained to generate a stand-alone question given a follow-up question and the preceding question-answer pairs. In the second stage, PR, the top-k relevant passages are retrieved from the index using BM25 using the rewritten question. Finally, in RC, a model is trained to extract an answer span from a passage or predict if the passage is irrelevant. The scores obtained from PR and RC are then combined as a weighted sum to produce the final score. The span with the highest score is chosen as the final answer.

7.1 Question Rewriting

We evaluate a co-reference model and several generative models on the QR subtask using the question rewrites in QReCC and the set of QR metrics selected in Section 6. The best performing model is then used in a combination with BERTserini to set the baseline results for the end-to-end QA task. All our Transformer-based models were initialized with the pretrained weights of GPT-2 (English medium-size) (Radford et al., 2019) and further fine-tuned on question rewrites from the QReCC training set (see Appendix A.1).

AllenAI Coref is the state-of-the-art model for coreference resolution task (Lee et al., 2018). We adapt it for QR with a heuristic that substitutes all coreference mentions with the corresponding antecedents from the cluster.

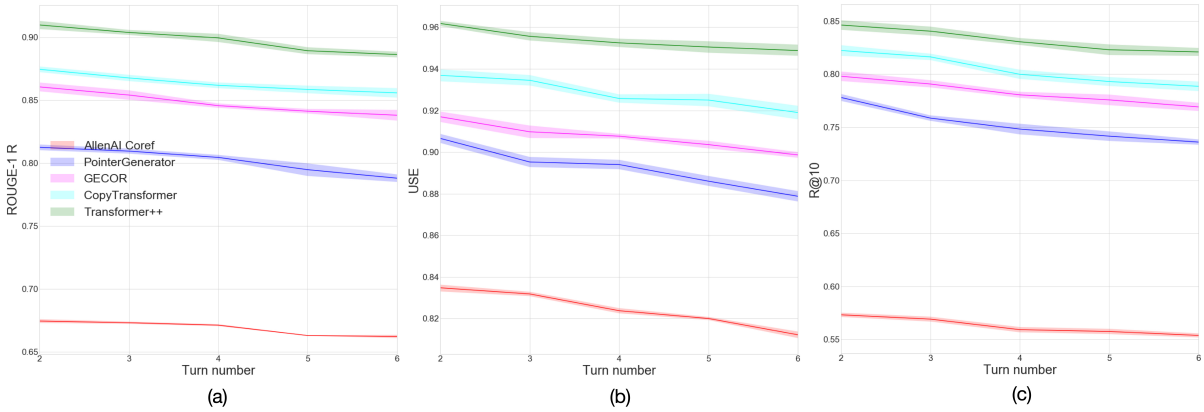


Figure 3: Rouge-1R, USE and R@10 metrics of baseline co-reference model, top-3 encoder-decoder models, and Transformer++ model based on dialogue turn number

PointerGenerator uses a bi-LSTM encoder and a pointer-generator decoder, which allows to copy and generate tokens (Elgohary et al., 2019).

GECOR uses two bi-GRU encoders, one for user utterance and other for dialogue context, and a pointer-generator decoder previously proposed for task-oriented dialogues (Quan et al., 2019).

Generator is a Transformer decoder model with a language modeling head (linear layer in the size of the vocabulary) (Radford et al., 2019).

Generator + Multiple-choice model has a second head for the auxiliary classification task that distinguishes between the correct rewrite and several noisy rewrites as negative samples (inspired by TransferTransfo (Wolf et al., 2019b)).

CopyTransformer uses one of the attention heads of the Transformer as a pointer to copy tokens from the input sequence directly (Gehrmann et al., 2018).

Transformer++ model has two language modeling heads that produce separate vocabulary distributions, which are then combined via a parameterized weighted sum (the coefficients are produced by combining the output of the first attention head and the input embeddings).

7.2 BERTserini

We implemented BERTserini following Yang et al. (2019) We use the standard BM25 ranking for passage retrieval with $k_1 = 0.82$, $b = 0.68$, which was previously found to work well for passage retrieval on MS MARCO. We then retrieve the top-100 relevant passages per question. Afterwards, we use BERT-Large fine-tuned for the task of reading comprehension. This model takes a question and each of the relevant passages as input and produces the answer span (Wolf et al., 2019a). BERT-Large produces a score (S_{BERT}), which is combined with the retrieval score for each of the passages (S_{Anserini}) through simple linear interpolation:

$$S = (1 - \mu) \cdot S_{\text{Anserini}} + \mu \cdot S_{\text{BERT}}$$

We pick the span with the highest score S as the answer. The parameter $\mu \in [0, 1]$ was tuned using a 10% random subset of the QReCC training set withheld from the BERT-Large training (we found $\mu = 0.7$ to work best).

BERT-Large was trained on human rewrites from the QReCC training set, and evaluated on the test set using either the original questions, human rewrites or the rewrites produced by Transformer++. The model is trained to either predict an answer span or predict that the passage does not contain an answer. "No answer" for the question is predicted only when neither of the relevant passages predicts an answer span. The model was trained on 480K paragraphs that contain the correct answers and 5K of other paragraphs as negative samples (see Appendix A.3 for more details).

8 Baseline Results

We use the results of QR to select the best model and then use it for the end-to-end QA task. Question rewrites are used as input for both passage retrieval and reading comprehension tasks. The effectiveness of the QR component is compared with the end-to-end model conditioned on the conversational context.

Table 5: Mean reciprocal rank, recall@10, and recall@100 for passage retrieval on test set questions.

Rewrite Type	MRR	R@10	R@100
Original	0.0343	6.12	11.71
Transformer++	0.1586	26.52	41.51
Human	0.1994	32.78	49.36

8.1 Question Rewriting Effectiveness

We analyze the effectiveness of our QR models by doing a 5-fold cross validation and obtaining the best performing metrics. Figure 3 contains 3 plots showing ROUGE 1-R, USE and R@10 across 5 turns. We start with the second turn because the first turn always is a self-contained query. The metrics across turns also stay stable with the same result for all the models. The Transformer++ model is stable with little variance in terms of its maximum and minimum metric values across all the best performing metrics.

Our evaluation results are summarized in Table 4. All generative models outperform the state-of-the-art coreference resolution model (AllenAI Coref). We noticed that PointerGenerator which employs a bi-LSTM encoder with a copy and generate mechanism outperforms Generator using Transformer alone. We could not find evidence that pretraining with an auxiliary regression task can improve the QR model effectiveness (Generator + Multiple-choice). Use of two separate bi-GRU encoders for the query and conversation context further improved the QR effectiveness (GECOR). Modeling both copying and generating the tokens from the input sequence employing the Transformer helped improve the effectiveness of the QR model (Copy-Transformer) compared to other existing generative models. Finally, obtaining the final distribution by computing token probabilities and weighting question and context vocabulary distributions with those probabilities helped improve over the best performing generative model (Transformer++).

8.2 Question Answering Effectiveness

Table 5 shows the mean reciprocal rank (MRR), R@10, and R@100 of using the original, Transformer++, and human rewritten questions. R@k is averaged across all questions. For a question, if R@k is 1.0, it means that there is a passage in the top-k at any rank such that the passage is relevant; and 0.0 otherwise. Table 6 shows the standard F1 and Exact Match metrics for extractive QA for each type of input question. In the "End-to-End" setting, the retrieval score was combined with the BERT reader score to determine the final span. In the "Known Context" setting, we use the relevant passage from the web page indicated by the human annotator, i.e., without passage retrieval. In the "Extractive Upper Bound" setting, we use a heuristic to find the answer span with the highest F1 score among the top-100 retrieved passages with human rewrite. This setup indicates the best the reader can do given the retrieval results.

Table 6: Mean F1 and Exact Match scores (%) on passages for extractive QA. "Known Context" assumes perfect retrieval. The "Extractive Upper Bound" assumes perfect single document span extraction.

Setting	Rewrite Type	F1	EM
End-to-End	Original	9.07	0.32
	Transformer++	19.10	1.01
	Human	21.82	1.23
Known Context	Original	17.24	1.90
	Transformer++	32.34	4.04
	Human	36.42	4.70
Extractive Upper Bound		75.45	25.07

The upper bound on the answer span extraction (F1 = 75.45) highlights the need for more sophisticated QA techniques than the standard reading comprehension approaches can offer now. Some answer texts in QReCC were paraphrased or summarised using multiple passages from the same web page. Abstractive approaches to answer generation are necessary to close this gap.

Even using single document span extraction techniques, there is a large room for improvement. Comparing "Known Context" to "End-to-End" we see losses introduced by the retrieval step, and comparing the "Extractive Upper Bound" to "Known Context" we see the sizeable margin of improvement available even for extractive models. This shows that even with competitive baselines the QA tasks are all far from solved.

In both Table 5 and 6 we see that human rewritten questions more than double the effectiveness of using original questions. In the absence of human rewritten questions, using Transfomer++ elevates the effectiveness of the QA tasks, getting it much closer to that proffered by human-level QR.

9 Conclusion

We introduced the QReCC dataset for open-domain conversational QA. QReCC is the first dataset to

cover all the subtasks relevant for conversational QA, which include question rewriting, passage retrieval and reading comprehension. We also set the first end-to-end baseline results for QReCC by evaluating an open-domain QA model in combination with a QR model. We presented a systematic comparison of existing automatic evaluation metrics on assessing the quality of question rewrites and show the metrics that best proxy human judgement. Our empirical evaluation shows that QR provides an effective solution for resolving both ellipsis and co-reference that allows to use existing non-conversational QA models in a conversational dialogue setting. Our end-to-end baselines achieve an F1 score of 19.10, well beneath the 75.45 extractive upper bound, suggesting not only room for improvement in extractive conversational QA, but that more sophisticated abstractive techniques are required to successfully solve QReCC.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset. In *Advances in Neural Information Processing Systems, 2016, NIPS 2016 3-8 December, Barcelona, Spain*.

Bill Byrne, Karthik Krishnamoorthi, Sankar Chinnadhurai, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kim Kyu-Young. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, page 4516–4525.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2019. Cast 2019: The conversational assistance track overview. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC*, pages 13–15.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5920–5926.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 687–692.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Query reformulation using query history for passage retrieval in conversational search. *CoRR*, abs/2005.02230.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 220–227.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *Association for Computational Linguistics*, pages 280–290.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *SIGIR 2020: 43rd international ACM SIGIR conference on Research and Development in Information Retrieval*.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4547–4557.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. Coqa: A conversational question answering challenge. *TACL*, 7:249–266.

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten de Rijke. 2020. Conversations with search engines.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2087–2097.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. Question rewriting for conversational question answering. *CoRR*, abs/2004.14652.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *SIGIR 2020: 43rd international ACM SIGIR conference on Research and Development in Information Retrieval*.

Nick Webb, editor. 2006. *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*. Association for Computational Linguistics, New York, NY, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019b. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77.

A Reproducibility

A.1 Training Transformer++ for Question Rewriting

Details about training setup of Transformer++ for question rewriting task is provided in Table 7. The Transformer head is initialized with the pretrained weights of GPT-2 (medium) and further fine-tuned on the QReCC train set. We use PyTorch implementation from HuggingFace.[3] Transformer++ is trained using model parallelism on 5 Tesla V100 GPUs with hyperparameter search trial.

A.2 Building Document Collection

Here we provide further details for building the document collection. If the web page of the provenance link containing the answer was not archived by the Wayback Machine yet, we trigger the archiving through the Wayback Machine API whenever possible. Overall, 2% of the annotated web pages could not be archived by the Wayback Machine due to the restricted access (such as the Quora website).

For the Common Crawl data, we take the index files from November 2019 and filter URLs to only those that are retrieved with HTTP status code 200 and those that are identified as English. We extract the pages from the Common Crawl WET files that correspond to these filtered URLs, and sample the first link out of every 100 links in each filtered WET file.

Overall, we find that 97.8% of unique web pages found by human annotators to contain answer and has an associated archived copy on the Wayback Machine. The final collection consists of both these pages from the Wayback Machine and random web pages from the Common Crawl.

After downloading the pages we extract all text from the page using the Beautiful Soup library.[4] We iterate through the web page by newlines, and accumulate the tokens for every line. Whenever the number of tokens reaches 220 or more, we emit a paragraph, and reset the token counter to 0. Note the last paragraph on the page may have fewer than 220 tokens. After segmentation, we have a total of 54,241,550 passages which we index using Pyserini 0.10.0.1. Hence we treat each passage as a single document.

[3] https://github.com/huggingface/transformers

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

A.3 Training BERT-L for Reading Comprehension

Below we provide details about the training setup for BERT-L used of the reading comprehension task in our experiments, which is similar to the extractive reader setup in Longpre et al. (2019) but using BERT-L. We train on the full data of the QReCC training set, using Human rewritten questions. Our implementation of the BERT question answering modules follows that of the standard PyTorch (Paszke et al., 2019) implementations from HuggingFace, and are trained on 4 NVIDIA Tesla V100 GPUs. The model is trained to predict an answer span or abstain if the passage has "No Answer". For every query we obtain up to 25 paragraphs from the document that contains the gold answer as identified by a human grader. The paragraph with the answer is always used for training, and a portion of the other paragraphs are used in training as No Answer or "negative" examples. Using the development set we tune several hyperparameters, most importantly the percentage of negative examples to retain for training ("Pct Neg. Ratio"). Fixed parameters and tuning details are shown in Table 8.

B Annotation Guidelines

Instructions for question rewriting:

- Rewritten questions should be as close to the original as possible.
- Questions should not contain any references to the previous context of the conversation.
- Avoid using any pronouns in question rewrites.

Instructions for answering questions:

- Put the rewritten question (original question if it is already self-contained) in a web search engine to produce the correct answer.
- Produce an answer, which should be short and brief with minimum information required to answer the question.
- The answers should be grammatically correct, do not contain special symbols or any additional mark-up.
- Produce an answer that would be most natural for a human conversation.

Table 7: Hyperparameter selection and tuning ranges for TRANSFORMER++ used for question rewriting.

MODEL PARAMETERS	VALUE/RANGE
Fixed Parameters	
Batch Size	16
Optimizer	Adam
Vocabulary Size	150,263
Transformer Head	GPT-2 (medium)
Learning Rate Schedule	Exponential Decay
Output Attention	True
Max Input Sequence Length	1024
Max Output Sequence Length	30
Num Hyperparameter Search Trials	500
Tuned Parameters	
Num Epochs	[50, 100]
Initializer Range	[0.01, 0.1]
Dropout	[0.05, 0.2]
Attention Dropout	[0.05, 0.1]
Residual Dropout	[0.05, 0.1]
Learning Rate	$[1e-3, 1e-1]$
Decay Steps	[6000, 10000]
Decay Rate	[0.7 , 0.9]
Activation Functions	[ReLU, Leaky ReLU, GELU]
General	
Model Size (# params)	350M
Avg. Train Time (per epoch)	12 hours

Table 8: Hyperparameter selection and tuning ranges for BERT-L used for reading comprehension.

MODEL PARAMETERS	VALUE/RANGE
Fixed Parameters	
Batch Size	32
Optimizer	Adam
Learning Rate Schedule	Exponential Decay
Num Epochs	2
Max Input Sequence Length	512
Max Span Length	30
Num Hyperparameter Search Trials	32
Tuned Parameters	
Learning Rate	$[1e-5, 5e-5]$
Pct Neg. Ratio	[0.01, 0.5]
General	
Model Size (# params)	330M
Avg. Train Time (per epoch)	8 hours

- Answers can contain up to a maximum of 30 words.
- When the answer is not a text, provide the source URL only (e.g., a geo-location on a map or a music video link).

C Pitfalls of the Query Rewriting Metrics

Our evaluation results show that the text similarity metrics, such as ROUGE and USE, often fall short to reflect semantic similarity in case of lexical paraphrases. Retrieval-based metrics, such as Recall@10, are able to demonstrate better correlation with human judgement. However, retrieval-based metrics are more expensive to compute since it requires an API call for every query. Also, they rely on the underlying collection as well as the ability of the search engine to handle paraphrases. Our experiments show that text similarity metrics, however flawed, are still able to provide a good proxy for quickly assessing QR performance and are suitable for comparing models in the development phase during parameter tuning. Retrieval-based metrics are useful to better approximate human judgement but can be computed for the best models only that were pre-selected using text similarity metrics.

ROUGE-1 R metrics provides a very rough estimate of the model performance by counting the number of words missing from the generated question rewrite in comparison with the ground truth rewrite and does not have any mechanism to distinguish which words are more crucial than others. As a result, a question missing only a single letter will receive the same score as a question missing one of its most informative words. For example, ROUGE("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday") = ROUGE("When did Gabriel Garcia die", "When did Gabriel Garcia *Marquez* die") = 0.75.

USE is more sensitive to such variations and can better pick up on the character-level similarities: compare to USE("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday")=0.96 and USE("When did Gabriel Garcia die", "When did Gabriel Garcia *Marquez* die") = 0.91.

Web search results, while most accurately correlates with human judgment, also reflect sensitivity of the retrieval algorithm to the query formulation as well as the collection-specific selectivity of the query terms. The resulting scores for our sample rewrites are R@10("When is Robert Downey *Jr* birthday", "When is Robert Downey *Jrs* birthday")=0.6 and R@10("When did Gabriel Garcia die", "when did Gabriel Garcia *Marquez* die") = 0.78.

D Examples of Query Rewrites

In Table 9 we show sample question rewrites from top 3 QR models along with conversational context.

E Examples of Answers Found

In Table 10 we provide two sample answers found by the baseline model. In the first example, the baseline system picked the same passage as the human annotator, but extracted a different answer span from this passage. In the second example, the baseline system picked a different passage than the human annotator. While the produced answers are not exact matches to the corresponding human answers, we consider them as correct upon manual inspection. This observation that a single question in QReCC can have multiple correct answers poses an important challenge for standard approaches to QA evaluation.

Table 9: A sample of conversation snippets from the QReCC test set with the question rewrites produced by Transformer++, CopyTransformer, and GECOR models.

Conversational context	**Transformer++**	**CopyTransformer**	**GECOR**
Q1: Did Nadia Comăneci win any Gold medals in the Olympics? *A1:* Nadia Comăneci is a five-time Olympic gold medalist. *Q2:* What about Silver?	Did Nadia Comăneci win any Silver medals?	Did Nadia win Silver medals?	What about Silver medals?
Q1: What is range in statistics? *A1:* The Range is the difference between the lowest and highest values. *Q2:* Describe some different metrics to interpret it.	Describe some different metrics to interpret range in statistics.	Describe some different metrics to interpret range in statistics.	Describe some different metrics to interpret range.
Q1: What is nominal GDP? *A1:* Nominal GDP is GDP evaluated at current market prices and includes all of the changes in market prices that have occurred during the current year due to inflation or deflation. *Q2:* What is the difference with real?	What is the difference between nominal and real GDP?	What is the difference with real GDP?	What is the difference with real GDP?
Q1: Tell me about lavender plants? *A1:* Lavandula is a genus of 47 known species of flowering plants in the mint family, Lamiaceae. It is native to the Old World and is found from Cape Verde and the Canary Islands, Europe across to northern and eastern Africa, the Mediterranean, southwest Asia to southeast India. *Q2:* What are the different types?	What are the different types of lavender plants?	What are the different types of plants?	What are the different types of plants?

Table 10: A sample of answers produced by our end-to-end baseline for conversational QA. The baseline model can also produce relevant answers using spans that differ from the answers provided by the human annotators.

Human re-written question	What are the educational requirements required to become a physician's assistant?
URL	https://www.geteducated.com/careers/how-to-become-a-physician-assistant
Predicted URL	https://www.geteducated.com/careers/how-to-become-a-physician-assistant
Human passage	. . . In most cases, a physician assistant will need a master's degree from an accredited institution (two years of post-graduate education after completing a four-year degree). . . . Most applicants to PA education programs will not only have four years of education, they will also have at least a year of medical experience. . . . five steps to becoming a PA: Complete your bachelor's degree (a science or healthcare related major is usually best); Gain experience either working or volunteering in a healthcare setting; Apply to ARC-PA accredited programs; Complete a 2-3 year, master's level program; Pass the PANCE licensing exam.
Found passage	(Same as human passage.)
Human answer	Complete your bachelor's degree (a science or healthcare related major is usually best); Gain experience either working or volunteering in a healthcare setting; Apply to ARC-PA accredited physician assistant programs; Complete a 2-3 year, master's level PA program;
Baseline model answer	a physician assistant will need a master's degree from an accredited institution (two years of post-graduate education after completing a four-year
Answer F1	15.38
Human re-written question	What tools were used in the neolithic event?
URL	https://sciencing.com/list-neolithic-stone-tools-8252604.html
Predicted URL	https://stmuhistorymedia.org/neolithic-era-technology-advances-and-beginnings-of-agriculture
Human passage	. . . By the time the Neolithic came around, hand axes had fallen out of favor . . . scientists consider the creation of all these tools a sign of early human ingenuity. Scrapers Scrapers are one of the original stone tools, found everywhere where people settled, . . . Blades While a scraper can be used for cutting into an animal, a longer, thinner blade can be inserted deeper into a carcass, . . . Arrows and Spearheads Arrows and spearheads are a more sophisticated shape than simple scrapers and blades. . . . Axes The polished stone ax is considered one of the most important developments of the Neolithic era. . . . Adzes The adze is a woodworking tool. . . . Hammers and Chisels Chisels were made by attaching a sharp piece of stone to the end of a sturdy stick . . .
Found passage	. . . The Neolithic Age was a period in the development of human technology, beginning about 10,000 BCE, in some parts of the Middle East, and later in other parts of the world, and ending between 4,500 and 2,000 BCE. . . . Hunting also became much easier to accomplish with the introduction new of stone tools. The most common tools used were daggers and spear points, used for hunting, and hand axes, used for cutting up different meats, and scrappers, which were used to clean animal hides.
Human answer	Scrapers. Scrapers are one of the original stone tools, found everywhere where people settled, long before the Neolithic Age began. ...Blades. ...Arrows and Spearheads. ...Axes. ...Adzes. ...Hammers and Chisels.
Baseline model answer	The most common tools used were daggers and spear points, used for hunting, and hand axes
Answer F1	19.05

3. SCAI-QReCC Shared Task on Conversational Question Answering

Bibliographic Information

Svitlana Vakulenko, Johannes Kiesel, Maik Fröbe: SCAI-QReCC Shared Task on Conversational Question Answering. LREC 2022: 4913-4922.

SCAI-QReCC Shared Task on Conversational Question Answering

Svitlana Vakulenko[★♠], **Johannes Kiesel**[♦], **Maik Fröbe**[♥]
[★]Amazon, Spain, svvakul@amazon.com
[♦]Bauhaus-Universität Weimar, Germany, johannes.kiesel@uni-weimar.de
[♥]Martin-Luther-Universität Halle-Wittenberg, Germany, maik.froebe@informatik.uni-halle.de

Abstract

Search-Oriented Conversational AI (SCAI) is an established venue that regularly puts a spotlight upon the recent work advancing the field of conversational search. SCAI'21 was organised as an independent online event and featured a shared task on conversational question answering, on which this paper reports. The shared task featured three subtasks that correspond to three steps in conversational question answering: question rewriting, passage retrieval, and answer generation. This report discusses each subtask, but emphasizes the answer generation subtask as it attracted the most attention from the participants and we identified evaluation of answer correctness in the conversational settings as a major challenge and a current research gap. Alongside the automatic evaluation, we conducted two crowdsourcing experiments to collect annotations for answer plausibility and faithfulness. As a result of this shared task, the original conversational QA dataset used for evaluation was further extended with alternative correct answers produced by the participant systems.

Keywords: Conversational Systems, Question Answering

1. Introduction

Conversational Question Answering (QA) is a challenging task at the current research frontier (Qiu et al., 2021; Kim et al., 2021) important for developing conversational information retrieval (conversational search) systems (Anand et al., 2019). In conversational QA, a system is required to return a correct answer given a question and the previous conversation turns. Such questions are often ambiguous outside of the conversational context and require incorporating additional information contained in the previous conversation turns. Moreover, evaluating systems for conversational QA, especially for questions that require long generative answers drawn from multiple information sources, remains an open research problem in its own right (Voorhees, 2003; Krishna et al., 2021; Siblini et al., 2021; Li et al., 2021).

This paper provides an overview of the SCAI-QReCC 2021 shared task on conversational question answering. It reports on the extended conversational QA dataset employed in this task, participating systems and insights gained from their performance, and the challenges we faced when evaluating the submissions and our approach for dealing with those. Specifically, we designed and tested a set of guidelines to support evaluation of (conversational) QA models.

The shared task was built around the recently introduced QReCC dataset (Anantha et al., 2021). QReCC contains sequences of conversational questions paired with answers produced by human annotators. Such sequences imitate a dialogue session with follow-up questions asked by the user. The annotators had access to a web search engine and every answer is based on the content of a single web page, while several web pages may be used to answer different questions within the same sequence. These web pages were downloaded and chunked into passages. The task of the system is to retrieve information from this passage collection and produce the correct answer based on this information.

Since QReCC contains only one answer provided by the human annotators, our goal was not only to evaluate the current state-of-the-art but also to collect alternative correct answers for this dataset. While the ground truth provides a single correct answer per question, in practice more than one answer can be considered correct.

We received 29 runs, including the submissions made by four participating teams and the results produced by our three baseline models. Each of these runs contains answers to all questions of the QReCC test set. To assess these runs, we employed a range of automated metrics and arranged two crowdsourcing tasks.

Human assessment remains crucial for QA evaluation since automated measures may only assess similarity to the ground truth but the answer may be correct even when it differs from the ground truth. Our goal was to find such answers and add them to the QReCC dataset. Since it is unfeasible to manually evaluate all 17K answers we collected, we came up with a set of techniques to elicit a smaller subset that was more likely to contain alternative correct answers.

Another major challenge we faced was in judging answer correctness. All teams participating in the shared task used generative models to produce the answers by conditioning on several previously retrieved passages. However, it was previously shown that such models tend to deviate from the information provided in the passages (Krishna et al., 2021). While these generated

♠ Research conducted when the author was at the University of Amsterdam.

Conversational Question: "Do **NPs** or **PAs** make more?"

Rewritten Question: "Do **nurse practitioners** or **physician assistants** make more?"

Human Answer: "According to the Bureau of Labor and Statistics (BLS), the **2017** median pay for **Nurse Practitioners** is $110,930 per year, or $53.33 per hour. In comparison, 2017 median pay for **Physician Assistants** is $104,860 per year, or $50.41 per hour."

Generated Answer: "The average salary for a **nurse practitioner** is $65,000. The average salary for a **physician's assistant** is $30,000."

Extracted Answer: "The BLS reports that the median annual wage for **nurse practitioners** was $109,820 as of May **2019** , while the median annual wage for **physician assistants** reached $112,260 during the same month."

Figure 1: A sample from the SCAI-QReCC results with the original conversational question, disambiguated rewritten question, a ground-truth human answer, an unfaithful generated answer from one of the participating systems and an alternative correct answer extracted from the submitted passages using our approach.

answers may seem plausible to a human annotator, they may contain factual errors.

Considering the challenges mentioned above, we propose to evaluate the submitted answers in two stages: (1) answer plausibility: does it answer the question?; (2) answer faithfulness: does it follow from the evidence available from the passage collection?

Since the length of a single passage makes them unscrutinizable for a human evaluator, we employ a token-overlap-based heuristic to extract shorter spans from the retrieved passages. Consider the example provided in Figure 1. While the human answer is outdated and the generated answer is not factually correct, they both allow us to find similar answer spans within the retrieved passages that can be either used as evidence or as the extracted answers themselves.

The rest of the paper is structured as follows. Section 2 describes the subtasks, baselines and participating systems. Section 3 lists the metrics we used for automated evaluation and summarises their results. In Section 4 we provide a brief overview of our human evaluation procedure, which is then described in more detail in Sections 5-6 Finally, Section 7 concludes with a summary of our main findings, lessons learned and directions for future work.

2. Task Setup

This section describes the setup of the 2021 edition of the SCAI-QReCC2021 shared task in more detail, including the three subtasks relevant for conversational QA, our baselines, and participating systems.

2.1. Subtasks

Following the same setup as in QReCC, we decomposed the end-to-end conversational QA task into three subtasks that can be implemented and evaluated separately: (1) question rewriting – ability of the model to correctly interpret and reformulate the question into its unambiguous equivalent; (2) passage retrieval – ability of the model to locate information relevant for answering the question; and (3) answer generation – ability of the model to produce faithful and grammatically correct answers.

While the conversational QA task can be approached end-to-end using a single model, such decomposition is beneficial for several reasons: (1) it allows to reuse the same passage retrieval and answer generation components already tuned for non-conversational QA (Vakulenko et al., 2021); (2) it allows to efficiently scale retrieval to large document collections using sparse representation approaches (Luan et al., 2021); (3) it enables a more thorough evaluation providing insights into the bottlenecks and opportunities to improve the end-to-end performance (Vakulenko et al., 2020).

2.2. Baselines

We introduced three baseline models to serve as reference points:

- *Basic*. This baseline implements a naive approach for question answering: it submits the question as the answer. Though not intuitive at first glance, this baseline is surprisingly hard to beat as most QA metrics are based on token overlap between submitted and ground truth answer, and the ground truth answers naturally often share tokens with the respective questions.
- *Simple*. This baseline implements low-effort approaches for each subtask. For question rewriting, it submits the question without modification. For passage retrieval, it employs Pyserini[1] BM25 with $k1 = 0.82$ and $b = 0.68$ as in the QReCC paper (Anantha et al., 2021). For question answering, it submits the sentence from the retrieved passages that contains the most of the stemmed noun phrases of the question. The baseline's source code is available on Github.[2]
- *GPT3*. This baseline uses the OpenAI GPT3 API (Brown et al., 2020) with default parameters (see Appendix) for question answering. The answers for 16,451 conversational questions of the test set were generated in 90 minutes for 33 USD. As the model prompt, all preceding questions and answers were prefixed with "Q:" and "A:" respectively and then concatenated.

[1] https://github.com/castorini/pyserini

[2] https://github.com/scai-conf/SCAI-QReCC-21/tree/main/code/simple-baseline

2.3. Participating Systems

We encouraged participants to submit working software through the TIRA (Potthast et al., 2019) platform but also allowed for traditional run file submissions. Each participant received access to a dedicated virtual machine with full admin rights and access to the QReCC dataset and a pre-build Anserini (Yang et al., 2017) index to deploy their software submissions and improve upon the available Pyserini baseline. We deployed the Pyserini baseline as software submission in TIRA and made the code available to simplify adaptations and encourage reproducibility because software submissions in TIRA can be executed on new datasets in the future without adoption. Still, all participants submitted run files and no working software because the deadline of the SCAI-QReCC 2021 shared task was close to the deadline of the TREC 2021 CAsT track.
Overall, four teams submitted results to the shared task:

- *Rachael* (Gonçalo Raposo and Coheur, 2022) implemented a three-stage pipeline that rewrites the question with T5 (Raffel et al., 2020) and summarizes the top-10 passages retrieved with BM25 for the rewritten question using PEGASUS (Zhang et al., 2020a) to generate the answer. The T5 model for query rewriting is pre-trained[3] on the CANARD dataset (Elgohary et al., 2019) and uses the questions and answers of previous turns and the current question to make the current question context-independent. The rewritten query is used to retrieve the top-10 passages with BM25 implemented in Pyserini. Finally, the rewritten query and the top-10 passages are concatenated and fed to the abstractive summarizer PEGASUS fine-tuned on the official QReCC training set. The resulting summary is returned as the answer.

- *Rali-QA* used the human rewritten questions to retrieve passages with a pipeline identical to our simple baseline (BM25 with $k1 = 0.82$ and $b = 0.68$) and a fine-tuned extractive BERT model for question answering on the top-k passages. The BERT model was fine-tuned for span extraction during a single epoch on the official QReCC training dataset, starting with the weights from a BERT model pre-trained on SQuAD v1.

- *Torch* uses a three-stage pipeline that rewrites the question with GPT2 (Radford et al., 2019), retrieves passages with BM25 and a BERT-based re-ranker, and generates the answer from the top-scored passage using T5 (Raffel et al., 2020). The question rewriting follows the idea of Yu et al. (Yu et al., 2020) and uses GPT2 fine-tuned on the official QReCC training set with the questions and answers of previous turns and the current question to rewrite the current question. The passage retrieval re-ranks the top-1000 results for the rewritten query of BM25 implemented in Anserini (Yang et al., 2017) ($k1 = 0.9$ and $b = 0.4$) using the OpenMatch (Liu et al., 2021b) BERT re-ranker pre-trained on MS-MARCO.[4] Initially, the answer generation was intended in two stages using two T5 models. The first stage was intended to use a T5 model to generate a dedicated answer for each of the top-10 passages from the BERT re-ranking by concatenating the passage to the rewritten question. The second stage was intended to use another T5 model that uses the ten generated answers as input to generate the final answer as output. Due to the length of the passages, the fine-tuning of the second T5 model failed. Hence, the final answer was generated using the top-passage concatenated to the rewritten question (leaving the second stage of answer generation for future work).

- *Ultron* use a two-stage pipeline rewriting the question with the sequence-to-sequence model BART (Lewis et al., 2020a) and generating the answers using RAG (Lewis et al., 2020b). The BART model uses the current question and the queries of the previous turns as input to rewrite the current question. The BART model was fine-tuned for question rewriting using the official QReCC training set. Due to the large size of the collection, team ultron tested three different indexes for answer generation with RAG: (1) the Wikipedia index, (2) a filtered version of the QReCC passages using the top-10 passages for the questions of all turns of the conversation retrieved with BM25, and (3) a filtered version of the QReCC passages using the top-100 passages for the questions of all turns of the conversation retrieved with BM25.

3. Automatic Evaluation

This section lists the metrics we used for automated evaluation and summarises their results.

3.1. Metrics

The following section describes the metrics employed for each of the three subtasks. All metrics are averaged over all turns in the test set. Our script to calculate these metrics is made available as open-source.[5]

Question Rewriting We employ the measure which achieved in the dataset paper the highest correlation with human judgements for this subtask, ROUGE-1 (Anantha et al., 2021). The question rewriting subtask

[3] https://huggingface.co/castorini/t5-base-canard

[4] https://huggingface.co/castorini/monobert-large-msmarco

[5] https://github.com/scai-conf/SCAI-QReCC-21/tree/main/code/evaluation-script. Due to the large model size, KPQA was computed using the original script in https://github.com/hwanheelee1993/KPQA

is only performed and evaluated on the *original* dataset as the questions in the *rewritten* dataset are already the ground-truth rewrites.

- *QR*. We report, ROUGE-1 (Lin, 2004) here, which is the unigram recall (on token level) between the automatically rewritten questions and the ground-truth one as recommended based on the experimental evaluation performed by (Anantha et al., 2021) (Section 6: Question Rewriting Metrics Validation).

Passage Retrieval From the dataset paper we adopt the use of the mean reciprocal rank to evaluate the ranking of passages retrieved for each question (Anantha et al., 2021). Although the use of MRR faces some criticism (Fuhr, 2017; Zobel and Rashidi, 2020), it is still used for evaluations in the TREC 2021 Conversational Assistance and Deep Learning tracks. We employ it here for comparability with previous work.

- *MRR*. This metric is equal to $1/r$, where r is the rank of the highest-ranked relevant passage. We employ the token overlap heuristic of Anantha et al. (2021) to determine relevance. Using token overlap is similar in spirit to several of the question answering metrics described below.

Question Answering We experiment with eight different metrics for comparing the generated answers with the ground-truth answers:

- *EM*. The "Exact Match" is 1 if the answer is identical to the ground-truth after lowercasing, stemming, punctuation, and stopword removal. An Exact Match of 0 corresponds to disjoint pairs of text.

- *F1*. The F-Measure is the harmonic mean of precision and recall. These correspond here to the fraction of shared tokens between predicted and ground-truth answer among the tokens in the predicted or in the ground-truth answer, respectively.

- *R1*. ROUGE-1: the same as *QR* for question rewriting.

- *POSS*. The POSSCORE (Liu et al., 2021a) computes the cosine similarity of averaged word embeddings between predicted and ground-truth answers, but does so separately and weighted for tokens with specific part-of-speech tags and tokens with other tags. We use the default tag set: ADJ, ADV, VERB, PROPN, NOUN.

- *SAS*. Semantic Answer Similarity (Risch et al., 2021) uses a cross-encoder neural network, where both predicted and ground-truth answer are first concatenated with a separator token in between and then fed into a language model for similarity prediction.

- *BERT*. BERTScore (Zhang et al., 2020b) computes the token-wise F-Measure (see above) between predicted and ground-truth answers, but matches tokens based on the highest cosine similarity of the respective contextual BERT embeddings and uses this similarity instead of a binary exact match. Moreover, matches are weighed by the inverse document frequency of the matched token.

- *BKPQA*/*RKPQA*. BERTScore-KPQA and ROUGE-L-KPQA (Lee et al., 2021) are modified version of BERTScore (see there) and ROUGE-L (Lin, 2004) that weight each token by a predicted importance of the answer token with respect to the question. This importance score is predicted using a fully connected neural network that is trained on extractive question answering datasets.

3.2. Results

The upper part of Table 1 compares our baselines (Section 2.2) and the participating systems (Section 2.3) on the test split of the QReCC dataset. The bottom part shows the scores when directly using the human rewritten questions as input instead of the original ones.

Question Rewriting This subtask is evaluated on the original dataset only. The QR column shows the ROUGE-1 that the approaches reached. The simple baseline returns the question as-is and is outperformed by every run of the two teams that implemented their own rewriting approaches. Both teams achieved similar scores, with team Rachael having a slight advantage indicating that T5 might be a good starting point for further improving question rewriting.

Passage Retrieval The teams Rachael, Rali-QA, and Torch submitted results for the passage retrieval subtask, which we compared in terms of the mean reciprocal rank (MRR). Nearly all submitted runs improve upon the simple baseline, which uses the BM25-retrieval model with a standard parameter set. The best run, by team Rachael, reaches even a more than twice as high score. However, the run with the highest QR does not reach the highest MRR. Indeed, the Pearson correlation of a runs QR and MRR for team Rachael is -0.35, indicating a weak negative correlation. The scores for QR and MRR are similar across team Rachael's runs, though. In general, a considerably higher MRR is reached when using human rewritten questions, showing the necessity of the rewriting step. Figure 2 visualizes this performance increase. Surprisingly, the simple baseline performed best in this case. A possible explanation is that the not-rewritten questions it uses provide a much worse starting point for passage retrieval than the automatic rewritten questions of the participating teams.

Question Answering All four teams participated in the question answering subtask. For this task, the baselines are consistently beaten by at least one participant

Team	Run	QR	MRR	EM	F1	R1	POSS	SAS	BERT	BKPQA	RKPQA
Original questions											
-	Basic baseline	-	-	0.000	0.114	0.095	1.283	0.207	0.422	0.432	0.064
-	GPT3 baseline	-	-	0.001	0.149	0.148	1.305	0.264	0.448	0.467	0.134
-	Simple baseline	0.571	0.065	0.001	0.067	0.150	1.490	0.162	0.367	0.426	0.097
rachael	2021-09-04-10-38-07	-	0.056	0.002	0.138	0.193	**1.583**	0.163	0.410	0.476	0.135
rachael	2021-09-08-07-07-57	0.675	0.135	0.006	0.187	**0.226**	1.570	**0.277**	**0.452**	**0.498**	0.175
rachael	2021-09-08-07-09-57	0.682	0.128	0.006	0.186	0.226	1.558	0.269	0.448	0.494	0.175
rachael	2021-09-08-15-40-34	0.679	0.133	0.007	0.176	0.211	1.456	0.254	0.420	0.460	0.164
rachael	2021-09-08-21-49-44	0.681	0.130	0.008	0.177	0.211	1.461	0.246	0.422	0.462	0.167
rachael	2021-09-15-09-05-06	0.673	**0.158**	**0.011**	0.179	0.212	1.333	0.254	0.405	0.444	0.172
rachael	2021-09-15-09-06-44	0.681	0.150	0.010	0.179	0.211	1.369	0.249	0.408	0.449	0.169
rachael	2021-09-15-09-07-49	0.676	0.157	0.010	0.187	0.219	1.399	0.264	0.418	0.457	0.175
rachael	2021-09-15-09-08-40	**0.685**	0.149	0.010	**0.189**	0.222	1.458	0.259	0.428	0.470	**0.178**
torch	usi_T5_raw2	0.657	0.082	0.001	0.137	0.200	1.451	0.221	0.415	0.467	0.117
Human rewritten questions											
-	Basic baseline		-	0.000	0.224	0.205	1.555	0.351	0.517	0.472	0.132
-	Simple baseline		**0.398**	0.001	0.098	0.282	1.666	0.372	-	-	-
rachael	2021-09-04-10-39-42		0.359	0.011	0.267	0.331	**1.674**	0.398	0.534	0.562	0.258
rachael	2021-09-06-09-21-43		0.359	0.018	0.290	0.339	1.649	**0.430**	**0.549**	**0.570**	0.277
rachael	2021-09-15-19-36-31		0.385	**0.028**	**0.302**	**0.345**	1.618	0.420	0.544	0.566	**0.290**
rali-qa	2021-09-09-13-01-07		0.269	0.003	0.166	0.212	1.385	0.264	0.407	0.457	0.174
ultron	2021-09-04-17-28-07		-	0.001	0.183	0.186	1.357	0.301	0.463	0.457	0.121
ultron	2021-09-08-15-04-28		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220
ultron	2021-09-08-15-07-30		-	0.001	0.187	0.189	1.380	0.306	0.472	0.465	0.123
ultron	2021-09-08-15-08-00		-	0.004	0.247	0.236	1.597	0.379	0.536	0.525	0.177
ultron	bart-large_top1bm25		-	0.000	0.017	0.017	0.150	0.111	0.046	0.048	0.016
ultron	distilbart-xsum-12-1_top1bm25		-	0.000	0.019	0.020	0.170	0.113	0.050	0.054	0.018
ultron	distilbart-xsum-12-3_top1bm25		-	0.000	0.022	0.023	0.175	0.117	0.052	0.056	0.021
ultron	rag-bm25_100		-	0.004	0.247	0.236	1.597	0.379	0.536	0.525	0.177
ultron	rag-dpr-hard-neg-bm25-top10		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220
ultron	rag-ft-dpr-hard-neg-bm25_10		-	0.015	0.261	0.258	1.565	0.383	0.533	0.539	0.220

Table 1: Evaluation results on the original dataset (top) and when using the human rewritten questions as input. Metrics are described in Section 3.1. A "-" denotes that no output was submitted for the respective task or the evaluation code failed (for BERT, BKPQA, and RKPQA).

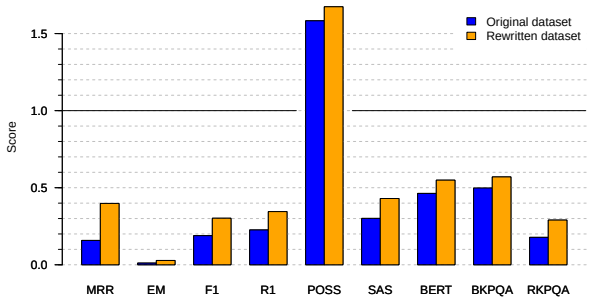


Figure 2: Highest scores reached per dataset variant for each metric. For all metrics except POSS a score of 1 corresponds to optimal performance.

for each metric. However, we find that the rankings of the different metrics often disagree, leaving otherwise an inconclusive image. As an extreme case, the runs by team Rachael that achieved the highest POSSCORE (for original and human rewritten questions) actually got the lowest score of their runs for most other metrics. The only exceptions are the BERTScore and BERTScore-KPQA metric when using original questions. Despite these differences, however, the metrics generally agree on ranking the runs of team Rachael high up, indicating the potential of their approach. Moreover, as Figure 2 shows, using human rewritten questions instead of the original ones does also improve question answering performance—for all metrics—, though to a lesser extent than for passage retrieval (MRR). To enrich these automated results and provide more insight, we also employed a human evaluation procedure outlined in the next section.

4. Human Evaluation

An overview of our human evaluation of the QA performance is given in Figure 3 and consists of two phases. We start with the set of candidate answers A_C for question q that was obtained from the runs submitted by the participants and our baseline approaches. Then, we use the SAS score (Risch et al., 2021) that shows similarity to the ground truth answer a from QReCC to filter

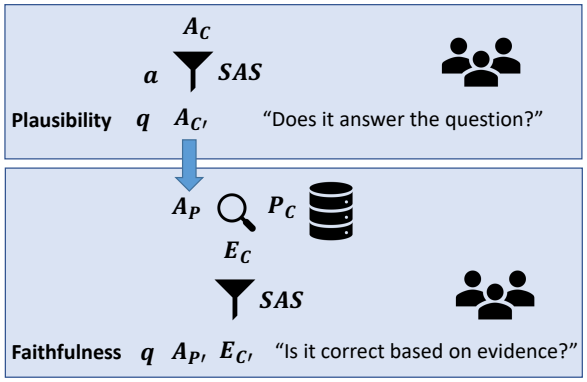


Figure 3: Our human evaluation including assessment of the answer plausibility and answer faithfulness.

out the candidate answers that are more likely to be correct. The subset with the highest SAS scores $A_{C'} \subseteq A_C$ is then used to crowdsource answer plausibility labels (see Section 5 for more details). The result of this phase is the subset $A_P \subseteq A_{C'}$ of all plausible answers identified for question q.

The next evaluation phase is designed to assess faithfulness of the plausible answers identified at the previous phase. For each plausible answer in A_P, we collect a set of evidence spans E_C by matching each answer to the top-k passages submitted by the participants P_C using a sliding window and a token-overlap heuristic. We keep the token-overlap threshold low to find semantic matches and then apply the SAS score to filter out the candidate spans. The resulting sets of spans extracted from the passages $E_{C'} \subseteq E_C$ paired with the corresponding plausible answers $A_{P'} \subseteq A_P$ is then used to crowdsource faithfulness labels (see Section 6 for more details).

5. Answer Plausibility

In this section, we describe our answer sampling approach and annotation instructions. We conclude with summarising our results.

5.1. Answer sampling

We obtained 16,736 answers, in total, from all the submitted runs including our baselines and participant systems. Our goal was to find alternative correct answers or answer variations that could help us to extend the ground-truth answers in QReCC. We also decided to focus on the questions on which the participating systems tend to disagree. To this end, we sampled at least four alternative answers per question submitted by the systems that are all distinct but at the same time semantically close to the ground truth answer using the SAS score (described in Section 3.1) with the threshold above 0.7. While the SAS score allows to find answer paraphrases beyond the lexical overlap matches, it also enables us to capture alternative answers that may contradict the ground truth answers (see the Generated Answer in Figure 1).

5.2. Annotation task

We asked the crowd workers using Amazon Mechanical Turk (MTurk) to judge the answers with respect to the questions (see Figure 4). This annotation task was estimated to take about 30 seconds for a single QA pair, on average. The workers were reimbursed with $0.06 per sample they annotated.[6]

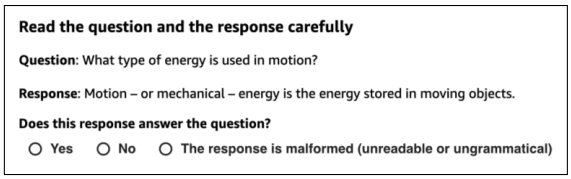


Figure 4: Task setup for answer plausibility annotation.

We started the crowdsourcing experiment with several smaller batches at first and performed the quality assurance of the results to select a pool of the trusted crowd workers who then completed the rest of the annotations. One of the authors did the quality assurance by manually examining the samples with disagreements. We collected two annotations per sample. The disagreements were manually resolved for the first batches. The rest of samples on which our pool of trust crowd workers disagreed were automatically discarded.

5.3. Results

After processing the annotations, we collected 1,863 answers judged as plausible, 108 malformed answers, and 107 implausible answers (see Table 2 for per run distribution). Thereby, we obtained 465 clusters with more than one plausible answers for the same question (avg: 4, max: 13 answers per cluster). In some cases those answers were paraphrases, shortened or more detailed versions of the same answer. In other cases, the answers are different, which may or may not be contradictory, such as pointing at different aspects in a definition (see the example provided in Figure 5).

6. Answer Faithfulness

This section describes our answer grounding approach that produces short evidence spans enabling the crowdworkers to efficiently judge answer faithfulness.

6.1. Evidence sampling

Evaluating answer correctness with respect to the passage collection (faithfulness) is a hard task because the passages are very long for crowd workers to go through. Our approach to sampling evidence spans is based on the observation that the human answers often paraphrase the original paragraph text in QReCC but,

[6]The rate was calculated based on an hourly wage of $7.25, which is the US federal minimum wage provided by the US Department of Labor.

Q: "What is a physician's assistant?"

A1: "A physician's assistant (PA) is a medical assistant who works under the supervision of a physician and is licensed to practice medicine in the state in which the patient resides."

A2: "A physician assistant is a person who has successfully completed an accredited education program for physician assistant, is licensed by the state and is practicing within the scope of that license."

A3: "A physician's assistant is a person who assists a physician in the performance of his or her duties."

A4: "A physician assistant is a medical professional who assists a doctor in the diagnosis and treatment of a patient."

Figure 5: A sample cluster of plausible answers for the same question from the SCAI-QReCC results.

in the majority of cases, those paraphrases stay sufficiently close to the original such that the token overlap heuristic helps to identify them.

Firstly, we use the same span heuristic proposed in (Anantha et al., 2021) to select short text spans within the passages with the maximum token overlap. We lower the token overlap threshold considerably to allow for non-verbal semantic matches as well. The evidence spans were trimmed to contain full sentences for readability. Secondly, we use the SAS score to compare between the matched sentences and all the generated answers to select a pool of evidence for each question. Similarly, we also sample additional evidence using the ground-truth answers from QReCC produced by human annotators.

Afterwards, we just merge all the evidence spans detected this way into a single paragraph. In this manner, we still attempt to evaluate the answers for cases when the collected spans may already contain sufficient evidence to judge the answer faithfulness even though they did not match any of the evidence spans directly.

6.2. Annotation task

We annotated two batches with 578 samples. Each sample contains a triple of a question, one of the plausible answers and the text, which is a concatenation of all the evidence spans obtained for this question (see Figure 6 for an example of one such sample). The workers were reimbursed $0.12 per sample since they had to read through the answers and the matched evidence spans as well. We followed the setup similar to the previous annotation task with two workers annotating the same sample. We also used the same pool of MTurk crowd workers that was selected during the previous annotation task. The annotation task distinguishes texts which are irrelevant or relevant to the question, and for relevant texts whether the answer can be deduced from the text ("contains information sufficient to judge the answer as correct"). Only triples with a relevant text and deducable answer are seen as faithful.

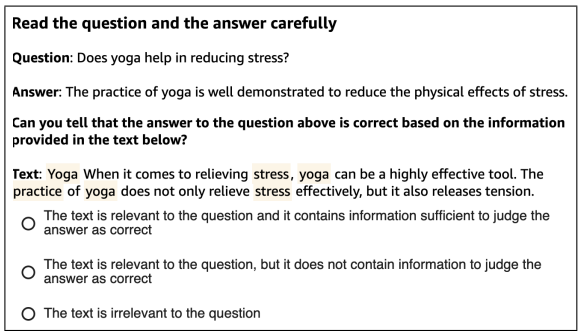


Figure 6: Task setup for answer grounding annotation with evidence sentences sampled from several retrieved passages by matching to the generated and ground-truth answers.

6.3. Results

After removing the samples on which the workers disagreed, we obtained 386 answers that were judged as faithful given the matched evidence spans (i.e., the first option in Figure 6). Our results are summarised in Table 2. Most of the answers were judged as faithful given the evidence spans we extracted, which shows the effectiveness of our evidence extraction approach and fidelity of the evaluated models. The table also demonstrates that some of the models that produced many plausible answers, such as GPT3, have a lower proportion of answers judged as faithful than other models. Note that this may also indicate that the answers generated by these models are very different from the retrieved passages. In this case, our approach is not sufficient to detect relevant evidence spans. Therefore, we believe that a reasonable requirement for generative QA models in the future shared tasks should be to provide the short evidence spans alongside with the passage IDs that can be used for answer evaluation.

7. Conclusion

Results of the SCAI-QReCC shared task identified main achievements as well as the major challenges when applying the state-of-the-art models to the task of open-doman QA. All of the submitted runs used a sparse index with BM25 for initial retrieval in combination with question rewriting for conversational QA. Due to the large collection size, none of the participant teams managed to scale dense passage retrieval that could allow to deploy an end-to-end conversational QA model. These results provide an important indicator of the technology maturity level for large-scale QA and conversational QA beyond Wikipedia-sized corpora.

Overall, we proposed and tested in practice an evaluation procedure that allowed us to compare the model

Team	Run	Question	Plausible	Implausible	Malformed	Faithful	Unfaithful
rachael	2021-09-04-10-39-42	rewritten	**183**	5	4	**37**	2
rachael	2021-09-08-21-49-44	original	133	6	4	30	1
rachael	2021-09-08-07-07-57	original	120	4	5	30	0
rachael	2021-09-15-09-07-49	original	103	4	6	29	1
-	GPT3 baseline	original	149	4	8	28	3
ultron	rag-bm25_100	rewritten	173	15	6	27	2
rachael	2021-09-06-09-21-43	rewritten	158	4	3	26	**4**
ultron	2021-09-08-15-04-28	rewritten	149	**16**	6	24	1
rachael	2021-09-15-19-36-31	rewritten	132	2	2	24	0
rachael	2021-09-15-09-06-44	original	73	0	4	22	1
rachael	2021-09-08-07-09-57	original	75	2	4	16	1
rali-qa	2021-09-09-13-01-07	rewritten	33	6	11	16	1
rachael	2021-09-08-15-40-34	original	41	6	2	14	3
torch	usi T5 raw2	original	36	7	**16**	14	0
ultron	2021-09-04-17-28-07	rewritten	117	13	7	13	0
rachael	2021-09-15-09-08-40	original	52	4	4	10	0
ultron	BART-large-top1BM25	rewritten	29	3	11	10	0
rachael	2021-09-15-09-05-06	original	52	2	1	9	0
rachael	2021-09-04-10-38-07	original	41	2	0	6	1
-	Simple baseline	rewritten	14	2	3	1	0
-	Simple baseline	original	0	0	1	0	0
Total			1863	107	108	386	21

Table 2: Human evaluation results of answer plausibility and faithfulness. The runs are ordered by the number of faithful answers. The highest values in each of the columns are highlighted in **bold**.

performance and discover new plausible and faithful answers. We used it to extend the original conversational QA dataset used for evaluation with multiple correct answers per question. While it is impossible to elicit a complete list of all correct answers for a given question, especially for an open-ended non-factoid question, this dataset is designed to improve our understanding of answer variations and specific properties important for a good answer. Our dataset is made available under the public URL: `https://doi.org/10.5281/zenodo.5749472`

Our evaluation results showed that modern QA models are able to produce fluent answers but we can not trust those answers to be correct. Since users are unaware of the models' limitations, they can be easily persuaded or misguided by plausible but unfaithful answers. This motivates the need to establish ways of producing high quality answers grounded in external information sources that can be referenced by the model.

We introduced an efficient approach for mining plausible and faithful answers. However, it is not suitable for evaluating answer faithfulness in cases where the generated answers are very different from the original text since it relies on the similarity measure between them. To make further progress on the generative QA task, models should be required to provide evidence alongside their answers explicitly indicating the answer provenance. Such evidence should be limited to relatively short spans of bounded length, such that they can be easily examined and assessed by human evaluators.

8. Bibliographical References

Anand, A., Cavedon, L., Joho, H., Sanderson, M., and Stein, B. (2019). Conversational search (dagstuhl seminar 19461). *Dagstuhl Reports*, 9(11):34–83.

Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., and Chappidi, S. (2021). Open-domain question answering goes conversational via question rewriting. In Kristina Toutanova, et al., editors, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, pages 520–534. Association for Computational Linguistics.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Hugo Larochelle, et al., editors, *33rd Annual Conference on Neural Information Processing Systems (NeurIPS'20)*.

Elgohary, A., Peskov, D., and Boyd-Graber, J. L. (2019). Can you unpack that? learning to rewrite questions-in-context. In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–

5923. Association for Computational Linguistics.

Fuhr, N. (2017). Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum*, 51(3):32–41.

Gonçalo Raposo, Rui Ribeiro, B. M. and Coheur, L. (2022). Question rewriting? Assessing its importance for conversational question answering. In *Advances in Information Retrieval. 44th European Conference on IR Research (ECIR 2022)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, March. Springer.

Kim, G., Kim, H., Park, J., and Kang, J. (2021). Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6130–6141. Association for Computational Linguistics.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4940–4957. Association for Computational Linguistics.

Lee, H., Yoon, S., Dernoncourt, F., Kim, D. S., Bui, T., Shin, J., and Jung, K. (2021). KPQA: A metric for generative question answering using keyphrase weights. In Kristina Toutanova, et al., editors, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, pages 2105–2115. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020a). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020b). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Li, H., Gao, T., Goenka, M., and Chen, D. (2021). Ditch the gold standard: Re-evaluating conversational question answering. *arXiv preprint arXiv:2112.08812*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Z., Zhou, K., Mao, J., and Wilson, M. L. (2021a). POSSCORE: A simple yet effective evaluation of conversational search with part of speech labelling. In Gianluca Demartini, et al., editors, *30th ACM International Conference on Information and Knowledge Management (CIKM'21)*, pages 1119–1129. ACM.

Liu, Z., Zhang, K., Xiong, C., Liu, Z., and Sun, M. (2021b). Openmatch: An open source library for neu-ir research. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2531–2535. ACM.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M. (2021). Sparse, dense, and attentional representations for text retrieval. *Trans. Assoc. Comput. Linguistics*, 9:329–345.

Potthast, M., Gollub, T., Wiegmann, M., and Stein, B. (2019). TIRA integrated research architecture. In Nicola Ferro et al., editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 123–160. Springer.

Qiu, M., Huang, X., Chen, C., Ji, F., Qu, C., Wei, W., Huang, J., and Zhang, Y. (2021). Reinforced history backtracking for conversational question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13718–13726. AAAI Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Risch, J., Möller, T., Gutsch, J., and Pietsch, M. (2021). Semantic answer similarity for evaluating question answering models. In *3rd Workshop on Machine Reading for Question Answering*, pages 149–157.

Siblini, W., Sayil, B., and Kessaci, Y. (2021). Towards a more robust evaluation for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6,*

2021, pages 1028–1034. Association for Computational Linguistics.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. (2020). A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, pages 7–16.

Vakulenko, S., Longpre, S., Tu, Z., and Anantha, R. (2021). Question rewriting for conversational question answering. In Liane Lewin-Eytan, et al., editors, *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 355–363. ACM.

Voorhees, E. M. (2003). Evaluating the evaluation: A case study using the TREC 2002 question answering track. In Marti A. Hearst et al., editors, *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics.

Yang, P., Fang, H., and Lin, J. (2017). Anserini: Enabling the use of lucene for information retrieval research. In Noriko Kando, et al., editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1253–1256. ACM.

Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P. N., Gao, J., and Liu, Z. (2020). Few-shot generative conversational query rewriting. In Jimmy Huang, et al., editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1933–1936. ACM.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations (ICLR'20)*. OpenReview.net.

Zobel, J. and Rashidi, L. (2020). Corpus bootstrapping for assessment of the properties of effectiveness measures. In Mathieu d'Aquin, et al., editors, *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1933–1952. ACM.

A. GPT3 Parameters

GPT3 baseline results in SCAI QReCC shared task were obtained using the following set of parameters in the official OpenAI API:

engine	curie
temperature	0
max tokens	100
top p	1
frequency penalty	0.0
presence penalty	0.0
stop	["\n"]

4. Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models

Bibliographic Information

Weijia Zhang, Jia-Hong Huang, **Svitlana Vakulenko**, Yumo Xu, Thilina Rajapakse, Evangelos Kanoulas: Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization Using Large Language Models. ICPR (19) 2024: 89-104.

Beyond Relevant Documents: A Knowledge-Intensive Approach for Query-Focused Summarization using Large Language Models

Weijia Zhang[1*], Jia-Hong Huang[1*], Svitlana Vakulenko[2], Yumo Xu[3], Thilina Rajapakse[1], and Evangelos Kanoulas[1]

University of Amsterdam[1], Amazon Alexa AI[2], University of Edinburgh[3]

Abstract. Query-focused summarization (QFS) is a fundamental task in natural language processing with broad applications, including search engines and report generation. However, traditional approaches assume the availability of relevant documents, which may not always hold in practical scenarios, especially in highly specialized topics. To address this limitation, we propose a novel knowledge-intensive approach that reframes QFS as a knowledge-intensive task setup. This approach comprises two main components: a retrieval module and a summarization controller. The retrieval module efficiently retrieves potentially relevant documents from a large-scale knowledge corpus based on the given textual query, eliminating the dependence on pre-existing document sets. The summarization controller seamlessly integrates a powerful large language model (LLM)-based summarizer with a carefully tailored prompt, ensuring the generated summary is comprehensive and relevant to the query. To assess the effectiveness of our approach, we create a new dataset, along with human-annotated relevance labels, to facilitate comprehensive evaluation covering both retrieval and summarization performance. Extensive experiments demonstrate the superior performance of our approach, particularly its ability to generate accurate summaries without relying on the availability of relevant documents initially. This underscores our method's versatility and practical applicability across diverse query scenarios.

Keywords: Query-focused summarization · Knowledge-intensive tasks · Large language models.

1 Introduction

Query-focused summarization (QFS) is a pivotal task with wide-ranging applications, spanning fields such as search engines and report generation [38].

*Equal contribution.

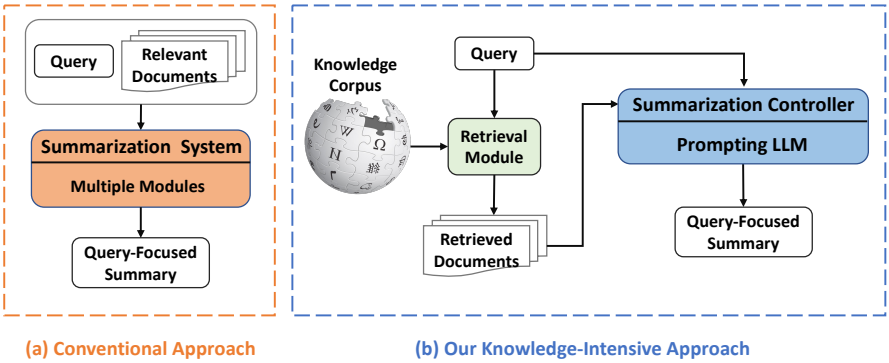


Fig. 1: The Comparison between (a) the conventional approach and (b) our knowledge-intensive approach. The conventional one assumes relevant documents are available. We aim to retrieve such documents from a large knowledge corpus.

It involves analyzing a textual query alongside a collection of relevant documents to automatically produce a textual summary closely aligned with the query [6,51,52,37]. This process aims to offer users relevant insights in a condensed format by applying information compression techniques to the provided document set [33,30,52]. The conventional approach to this task assumes the availability of a set of relevant documents and creates the summary based on the information within this document set. However, in practical scenarios, this assumption may not always hold.

Consider highly specialized or niche topics as an illustration. In fields such as advanced scientific research, specialized technology, or obscure hobbies, documentation may be limited or fragmented, particularly for cutting-edge or esoteric subjects. Similarly, think about future predictions or speculations. Queries related to future events, trends, or predictions may lack existing documents as they entail speculation or anticipation of events yet to unfold. For instance, questions about the outcome of upcoming elections, the impact of emerging technologies, or the trajectory of financial markets might not be documented until after these events have occurred. In the aforementioned practical scenarios, the traditional setup of QFS with compression techniques may not work well due to the scarcity or absence of relevant documents.

To tackle this challenge, our approach reframes QFS as a knowledge-intensive (KI) task setup [43]. Unlike the traditional setup where relevant documents are assumed to be readily available, our knowledge-intensive approach only assumes access to a large-scale knowledge corpus. Our proposed approach comprises a retrieval module and a summarization controller. Given a textual query as the initial input, the retrieval module returns potentially relevant documents from the knowledge corpus, eliminating the dependence on pre-existing document sets. These retrieved documents, along with the query, are then forwarded to the

summarization controller. The summarization controller harmoniously merges an advanced large language model (LLM)-powered summarizer with a carefully designed prompt, guaranteeing the produced summary is both comprehensive and directly addresses the query. To illustrate the aforementioned process, we make a comparison between the conventional approach and our proposed approach in Figure 1.

To validate the effectiveness of our proposed approach within the KI setup, we have introduced a specialized dataset by combining existing QFS datasets [6] with three various knowledge corpora. We also enhance the dataset with human-annotated relevance labels to facilitate comprehensive evaluation for both retrieval and summarization performance. This dataset serves as the foundation for a fair evaluation of our approach and baseline models. Extensive experiments conducted in this work confirm the effectiveness of our proposed method, surpassing baseline models and showcasing its superiority in generating query-focused summaries. Notably, our approach stands out for its ability to generate accurate summaries without relying on the availability of relevant documents at the initial input stage. This underscores the versatility of our method in addressing diverse query scenarios, further validating its practical applicability.

The primary contributions of this work can be summarized as follows:

1) **Introduction of a Knowledge-Intensive Approach**: We introduce a new knowledge-intensive approach for QFS, departing from the conventional paradigm that relies on the availability of relevant documents in the initial input stage. This innovative method involves the direct retrieval of pertinent documents from an extensive knowledge corpus, leveraging the provided textual query as the basis.
2) **Development of a Summarization Controller**: We propose a summarization controller that orchestrates the process of generating query-focused textual summaries using an integrated LLM-based summarizer with a specially designed prompt. This controller streamlines the summarization process, ensuring alignment with the query intent.
3) **Creation of a Specialized Dataset**: To assess the effectiveness of the proposed approach, we introduce a new dataset explicitly designed for knowledge-intensive QFS. This dataset, along with human-annotated relevance labels, serves as a standardized benchmark for evaluating both retrieval and summarization performance.
4) **Demonstration of Superior Performance**: We present extensive experimental results showcasing the superior performance of the proposed approach compared to baseline models. Particularly noteworthy is the approach's capability to generate accurate summaries without the prerequisite of available relevant documents at the initial input stage. This feature underscores the method's versatility and practical utility across a range of query scenarios.

2 Related Work

In this section, we cover three lines of research: query-focused summarization, knowledge-intensive language tasks, and open-domain summarization, all related to our study.

2.1 Query-Focused Summarization

Query-focused summarization aims to generate a summary tailored to a given query from a set of topic-related contexts. For instance, the contexts can be texts [6,37], tables [60,56], or videos [18,16,19,21,14,17]. We mainly focus on textual contexts. Conventional QFS methods primarily focused on sentence-level extraction techniques to eliminate redundant information [33,30,51,52]. However, they heavily rely on pre-existing relevant documents and are typically developed for a limited set of closed-domain documents (∼50 documents). In contrast to all previous work, our work does not assume the existence of relevant documents to the given query. Instead, our approach handles open-domain document-level retrieval, aiming to identify a small set of relevant documents from a pool of millions of candidates, accommodating more diverse topics and contexts.

Another line of research aims to automatically create larger QFS datasets to alleviate data scarcity, utilizing information retrieval or clustering methods [40,28,42,37].However, these datasets lack a standardized benchmark to assess the retrieval performance, as they lack relevance annotations for retrieved documents. In contrast, our dataset includes human annotations for the relevance of retrieved documents. This facilitates retrieval evaluation and provides insights into understanding the effects of retrieval errors on summarization performance.

2.2 Knowledge-Intensive Tasks

Knowledge-intensive language tasks [43] involve addressing user needs by leveraging a large-scale knowledge corpus. These tasks produce various types of outputs. For instance, they can generate short-form factual answers [4,15,22,20,63], dialog responses [7], or face-checking predictions [48]. One of the most related KI tasks is a long-form question answering (LFQA) [9,47,55], where the goal is to provide comprehensive answers to given questions. Compared to QFS, the key difference between LFQA and QFS lies in their objectives and the role of input documents. LFQA provides detailed answers to complex questions, whereas documents are optional external sources to enhance answer quality [9,47]. For instance, recent studies [1] investigated the effectiveness of LLMs in LFQA without relying on external documents. Additionally, generating long-form answers does not necessarily require multiple documents. In contrast, QFS requires generating a concise summary from multiple documents, tailored to a specific query. In this case, the relevant documents are essential inputs to generate the summary [51,30].

Furthermore, existing LFQA tasks either struggle with accurately grounding their answers in the provided documents [27] or they heavily rely on the inherent ambiguity present in questions [39]. In comparison, our work has two significant

advantages: 1) Our approach ensures generated query-focused summaries are firmly based on the content of the retrieved documents; and 2) The queries are unambiguous and accurately reflect complex information needs in QFS, resulting in naturally longer summaries to address various information needs adequately.

2.3 Previous Attempts on Open-Domain Summarization

Our preliminary research, first presented in the arXiv and non-archival workshop versions [57,58], marks the initial phase of investigating the complexities of open-domain QFS. These findings have yet to be officially published in any conference proceedings or archival workshops. Building upon this work, recent studies [12,62] have embarked on similar investigations. They either leverage multi-document summarization (MDS) datasets [8] or adapt query-based single-document summarization datasets [61,49]. In comparison, this work has two significant differences: 1) Previous studies lack human annotation for the relevance of the retrieved documents, raising concerns about the effectiveness of their evaluation of retrieval models. Instead, we conduct human annotations to ensure high-quality relevance annotations, thus enabling more accurate evaluation of retrieval models; and 2) Previous studies primarily focused on internal documents within specific domains, such as news or medicine, limiting their analysis to closed domains. In contrast, our study takes a broader approach by considering both internal documents and a significantly larger external knowledge corpus, e.g., Wikipedia. This broader scope enhances the applicability of our findings to real-world scenarios.

3 Knowledge-Intensive Approach

In this section, we formulate our knowledge-intensive task setup and detail our proposed approach.

3.1 Task Formulation

To better understand the difference between traditional QFS and our knowledge-intensive setup, we first describe the traditional QFS and then formulate our proposed knowledge-intensive setup.

Traditional QFS. Let the tuple $(q, \mathcal{D}, s)$ denotes an example in the traditional QFS task, where q is a given query, $\mathcal{D} = \{d_1, \ldots, d_m\}$ is a set of relevant documents and s is the gold summary for the document set $\mathcal{D}$ and the query q. The goal of QFS is to produce the query-focused summary given the documents and the query as input: $(q, \mathcal{D}) \rightarrow s$. It is worth noting that $\mathcal{D}$ is collected manually by human annotators and usually includes tens of documents.

Knowledge-Intensive QFS. In this work, we reframe the QFS task to a knowledge-intensive task setup as $q \rightarrow s$. As shown in Figure 1, we do not rely on the relevant document set $\mathcal{D}$. Instead, we assume the access to a large-scale knowledge corpus: $\mathcal{K} = \{d_1, \dots d_n\}$. In practice, $\mathcal{K}$ consists of millions of documents, i.e., $n \gg m$. It is worth noting that our goal remains to generate the query-focused summary s with respect to the query q. However, different from the traditional setup, there is no guarantee that the provided documents are relevant to the query. Instead, there are millions of irrelevant documents in the knowledge corpus. Thus, an effective information retrieval model is necessary to supplement the extended setup, which is described below.

3.2 Methodology

Our approach includes two components: a retrieval module and a summarization controller, as illustrated in Figure 1. The retrieval module returns top-k documents from the knowledge corpus based on the given query. Then the summarization controller takes the query and retrieved documents as the inputs to generate a query-focused summary.

Retrieval Module. Given a query q and a document d_i from the knowledge corpus $\mathcal{K}$, the retrieval module first estimates the relevance $Rel(q, d_i)$ between q and d_i and then ranks all documents based on their relevance scores. Typically, we only consider top-k retrieved documents. There are two main categories of retrieval models: sparse and dense models [59]. The former estimates relevance between the query and the document using weighted counts of their overlapping terms. The latter first transforms both the query and the document into vectors within an embedding space. The relevance is then computed using the dot product of their respective vectors. In this work, we explore representative sparse and dense models, which are BM25 [45] and Dense Passage Retrieval (DPR) [25], respectively.

Summarization Controller. After we obtain top-k retrieved documents from the retrieval module, we feed the query and the documents into the summarization controller to generate a query-focused summary. As large language models (LLMs), such as GPT-3.5 [41], have exhibited remarkable generation ability in many text generation tasks [29]. Inspired by Chain-of-thought prompting methods [50,26], we introduce a multi-step summarization controller. As shown in Prompt 3.1, the controller involves two primary steps: identifying query-relevant information and generating a controllable summary. First, to ensure the summary's relevance to the query, we explicitly instruct the LLM to identify query-relevant information within the documents. It is worth noting that the LLM can determine information at various levels, including phrases, sentences, and paragraphs. Subsequently, we instruct the LLM to write a controllable summary that meets the length limit of 250 words based on the query-relevant information. This ensures the summary is sufficiently comprehensive and relevant to the

query. Moreover, we include few-shot demonstrations in the prompt to enable in-context learning for the LLM, which has been proven to be effective in many natural language processing tasks [3].

Prompt 3.1: Summarization Controller

Instruction: You will be given a query and a set of documents. Your task is to generate an informative, fluent, and accurate query-focused summary. To do so, you should obtain a query-focused summary step by step.

Step 1: Query-Relevant Information Identification
In this step, you will be given a query and a set of documents. Your task is to find and identify query-relevant information from each document. This relevant information can be at any level, such as phrases, sentences, or paragraphs.

Step 2: Controllable Summarization
In this step, you should take the query and query-relevant information obtained from Step 1 as inputs. Your task is to summarize this information. The summary should be concise, include only non-redundant, query-relevant evidence, and be approximately 250 words long.

Demonstrations:
Few-shot human-written demonstrations.

Query: {*Input query*}
Documents: {*Retrieved documents*}

4 KI-QFS Dataset

In this section, we describe dataset collection and relevance annotation process for our knowledge-intensive setup. The dataset includes a collection of query-summary pairs and three alternatives of knowledge corpora. The relevance annotation aims to facilitate retrieval evaluation.

4.1 Dataset Collection

Collecting Query-Summary Pairs. We build our dataset on the top of query-summary pairs (q, s) on existing QFS resources. Specifically, we adopt DUC 2005-07 [6], three standard QFS benchmark datasets.[1] The DUC datasets consist of three subsets collected for the Document Understanding Conferences (DUC) from 2005 to 2007. Each subset contains 45-50 clusters. Each cluster (or a topic) contains a query, a set of topic-related documents, and multiple reference summaries. As relevant documents are inaccessible in our dataset, we only collect query-summary pairs first. The relevant documents will be used to build a knowledge corpus, which is described below. For data split, we follow previous

[1] We take DUC as an example, but nothing prohibits exploring other QFS resources.

work [34,30] to use the pairs of the first two years (2005-2006) as the training set and randomly select 10% from the training set as the validation set. We leave the third subset (2007) as the test set. Finally, the training set, validation set, and test set contain 90, 10, and 45 examples, respectively.

Building Knowledge Corpora. We explore three alternatives of knowledge corpora CORPUS for the query-summary pairs above. Specifically, we first follow the standard data processing for large-scale knowledge sources [25] to collect documents, where we split each origin document into non-overlapping context documents with 100 words maximum. [2] The knowledge corpora are as follows: 1) The first knowledge corpus, $\text{CORPUS}_{\text{Int}}$, is an internal collection, where we take all clusters of DUC documents from the three years to form this collection, which results in about 32K documents in total. As the queries have an explicit connection with the documents, this collection can be considered an in-domain knowledge corpus, where relevant documents are relatively easy to find. 2) However, as our main goal is to explore knowledge-intensive QFS on the large-scale knowledge corpus, we consider the second external knowledge corpus named $\text{CORPUS}_{\text{Ext}}$. We use the Wikipedia dump in the KILT benchmark [43] to form this corpus, resulting in about 21 million documents in total. 3) However, as $\text{CORPUS}_{\text{Ext}}$ is a Wiki-based corpus, there is no guarantee that the collection contains sufficient evidence to answer the query since the reference summary is derived from the content of the original DUC documents. To this end, we build an augmented knowledge corpus called $\text{CORPUS}_{\text{Aug}}$ by combining the previous two collections. We merge all the documents of $\text{CORPUS}_{\text{Int}}$ and $\text{CORPUS}_{\text{Ext}}$ into this collection ensuring that summaries can be grounded in the collection and that the collection is large-scale to make the task challenging.

4.2 Relevance Annotation

As the retrieval process is necessary for our setup, it is important to evaluate the performance of retrieval models so we can better analyze their effect on summarization. However, the DUC datasets do not come with relevance labels to designate which document provides evidence for the final summary. A possible solution to that would be to do a lexical match between the individual summary sentences and the available document collection, however, such an automatic evaluation has its disadvantages (e.g. a summary sentence may appear in a document but outside the right context, lexical overlap may pay attention to insignificant words in the summary, and there may be a lexical gap between semantically similar sentences). Therefore, we collect human annotations through Amazon Mechanical Turk (MTurk) to create our evaluation set.

As the main goal of the annotation process is to obtain relevant documents that support the summary, we begin with filtering out irrelevant documents,

[2] Some studies [25] also use passages to name the processed text chunks, but we adhere to the term "documents" to maintain the coherence of the paper.

which are the majority of millions of candidate documents in the knowledge corpus. Following the TREC document retrieval task [5], we apply retrieval models to choose top-ranked documents, as discussed in subsection 3.2. We first perform both BM25 and DPR on $\text{CORPUS}_{\text{Int}}$ and $\text{CORPUS}_{\text{Ext}}$, and construct top-50 pools, which result in a maximum of 200 candidate documents for each query. As there are 45 queries in our test set, we end up with 7761 query-document pairs to be annotated. We then design a web annotation interface shown to MTurk workers, which contains a query, a reference summary, and at most five documents. We ask workers to label whether a document contains either part of the summary or evidence to a query, adapting the annotation instructions in [5]. The judgments are on a 4-point scale from 0 to 3: (i) **0-irrelevant**: the document has nothing to do with the query; (ii) **1-weakly related**: the document seems related to the query but fails to contain any evidence in the summary; (iii) **2-related**: the document provides unclear information related to the query, but human inference may be needed; and (iv) **3-relevant**: the document explicitly contains evidence that is part of the summary. Prior to the collection of the annotations we performed two pilot studies to improve our annotation interface and the instructions. We collect three annotations per document and use a majority vote to determine the final relevance label. The distribution of relevance labels is 837 relevant, 5811 related, 1097 weakly related, and 16 irrelevant documents. We also measure inter-annotator agreement (IAA) using Fleiss Kappa [10]. A Fleiss Kappa score of 0.42 was obtained, which indicates a moderate agreement.

5 Evaluation Metrics

In this section, we describe evaluation metrics. As our approach includes a retrieval module and a summarization controller, we evaluate both components using suitable evaluation metrics.

5.1 Retrieval Evaluation

To evaluate retrieval effectiveness, i.e., the ability of the retrieval module to identify and retrieve evidence to be used to compose the summary, we use standard information retrieval (IR) metrics, and in particular precision@k (P@k) and recall@k (R@k) where k denotes the cut-off. For P@k and R@k, we map the human-annotated judgment level 3 to positive and judgment levels 0-2 to negative and report the corresponding results for $k \in \{10, 50\}$.

5.2 Summarization Evaluation

Lexical Metrics. In our experiments, we measure the accuracy of generated summaries against gold summaries. Following previous work [30,52], we employ the commonly-used standard ROUGE-1, ROUGE-2, and ROUGE-SU4 [35], which evaluate the accuracy on uni-grams, on bi-grams, and on bi-grams with a maximum skip distance of 4, respectively. We report the F_1 score with a

maximum length of 250 words for the metrics. It is worth noting that there are multiple reference summaries on our dataset, we take the average of the ROUGE scores for all summaries as the final reported ROUGE score.

Semantic Metrics. However, ROUGE-based metrics only measure lexical overlap between generated summaries and gold summaries, failing to capture their semantic similarities. Recent semantic metrics, such as BERTScore [54] and BARTScore [53], have been shown to be effectively correlated with human judgments [23]. Therefore, we also report them in our experiments. Specifically, we report the F1 score of BERTscore and use recommended `deberta-xlarge-mnli` [13] as the backbone. For BARTScore, we use the recall version of BARTScore as it shows more effective performance in summarization tasks [53], where we use `bart-large-cnn` as the backbone. It is worth noting that the raw scores of BARTScore are negative log probabilities that are difficult to explain, we follow [46] to normalize them to positive scores with an exponential function.

6 Experimental Setup

In this section, we describe the baseline models for comparison in our experiments. We then discuss the implementation details.

6.1 Baselines

We compare our proposed method with the following two families of baseline models:

Weakly Supervised QFS Models. To illustrate new challenges brought by our knowledge-intensive setup, we first include two weakly supervised models designed for traditional QFS tasks: 1) QuerySum [51]: an extractive QFS model leveraging distant supervision signals from trained question answering (QA) models to select salient sentences as the summary. 2) MargeSum [52]: an abstractive QFS model that employs generative models trained on generic summarization resources to boost sentence ranking and summary generation, achieving competitive performance in the weakly supervised setup. We discuss the adaption of two models to our knowledge-intensive setup in the implementation details.

Supervised Retrieval-Augmented Models. We employ supervised Retrieval-Augmented Generation (RAG) models below: 1) RAG-Sequence [32]: a generative model in which document retrieval and summary generation are learned jointly. As RAG models have two variants including RAG-Token and RAG-Sequence, we mainly report the results of RAG-Sequence as it shows better performance in knowledge-intensive tasks. Specifically, the RAG-Sequence model employs DPR [25] for retrieval and $\text{BART}_{\text{large}}$ [31] for generation. 2) Fusion-in-Decoder (FiD) [24]: an encoder-decoder architecture which has achieved state-of-art

performance in knowledge-intensive tasks [2]. In this model, encoded representations for retrieved documents are first concatenated and then fed into the decoder to generate the output. Following the origin paper, we use $T5_{base}$ [44] as the base model.

LLM-based RAG Models. In addition to supervised RAG models, we compare our approach with LLM-based RAG models [11]. Specifically, we employ BM25 or DPR for retrieval and GPT-3.5 for generation, referring to this baseline as NaiveRAG. Following the prompting strategy from Gao et al. [11], we instruct GPT-3.5 to answer a given query based on the associated retrieved documents. Notably, NaiveRAG does not require fine-tuning.

6.2 Implementation Details

For retrieval models, we use the library Pyserini [36] to implement BM25 and DPR.[3] For the summarization controller, we employ GPT-3.5 using the OpenAI API service, where we use long-context version `gpt-3.5-turbo-0613`.[4] We utilize 3-shot demonstrations in few-shot settings and present results for both zero-shot and few-shot settings. The temperature and top-p are set to 0.1 and 0.95, respectively, with a maximum output length of 400 tokens. The number of top-k retrieved documents is fixed at 50. Notably, we employed the long-context version of GPT-3.5, which supports a maximum context window of 16,385 tokens. Given that each retrieved document contains up to 100 words, when we concatenate 50 documents into a prompt, the total length of the prompt is roughly $7,118$ tokens on average, which is comfortably below the maximum limit.

For QFS models, as they have achieved competitive performance on the QFS datasets in a weak supervision manner, we do not fine-tune them on our dataset. For inference, although they include a retrieval module that performs in-domain evidence estimation, their retrieval models are not designed to handle large-scale, open-domain knowledge bases, such as $\textsc{Corpus}_{\text{Ext}}$. Therefore, we first retrieve top-1000 documents using BM25 for each query and then feed them as inputs to their customized sentence extraction modules. We follow the original papers [51,52] to set their hyper-parameters.

For retrieval-augmented models, as they are originally designed for short-form open-domain QA [4]. For fair comparisons, we fine-tune them on our training set. We follow the original papers [32,24] to set their associated hyper-parameters and training setups. For inference, we set the maximum decoded length and beam size to 250 and 5, respectively. All experiments are conducted on a single A6000 GPU.

7 Results and Analysis

Our experimental results primarily address the following research questions (RQs):

[3] https://github.com/castorini/pyserini
[4] https://platform.openai.com

Corpus	Model	P@10	P@50	R@10	R@50
$\text{CORPUS}_{\text{Int}}$	BM25	**16.0***	**12.0**	**8.2***	**30.2***
	DPR	11.6	11.5	6.0	27.8
$\text{CORPUS}_{\text{Ext}}$	BM25	12.7	8.6	**7.0**	22.2
	DPR	**13.8***	**12.6***	6.9	**29.7***
$\text{CORPUS}_{\text{Aug}}$	BM25	12.2	8.9	6.8	23.2
	DPR	**14.4***	**12.5***	**7.4***	**30.3***

Table 1: Retrieval evaluation of different models on the test set of our dataset over three knowledge corpora. P@k and R@k denote precision and recall regarding cutoff k. The best results are in **bold**. *means that the improvement is statistically significant (a two-paired t-test with p-value < 0.01).

RQ1 What is the retrieval effectiveness of different retrieval models in the knowledge-intensive setup?
RQ2 How does our approach compare to baseline models in the knowledge-intensive setup?
RQ3 What is the effect of our knowledge-intensive setup compared to the traditional QFS?

7.1 Retrieval Results

To answer **RQ1**, we evaluate the retrieval effectiveness of BM25 and DPR on the test set of our dataset among three different knowledge corpora. The results are shown in Table 1. We find that BM25 outperforms DPR on $\text{CORPUS}_{\text{Int}}$. One plausible explanation is that the internal knowledge corpus ($\text{CORPUS}_{\text{Int}}$) comprises documents sourced from human-curated DUC datasets, potentially containing more relevant keywords. These keywords can be easily retrieved by BM25 which is a term-based IR method. Conversely, DPR exhibits better performance on $\text{CORPUS}_{\text{Ext}}$. This is because it is pre-trained on Wikipedia data dumps, allowing it to better capture semantic representations of Wikipedia documents, thus enhancing retrieval performance. Moreover, we observe DPR's superior performance over BM25 on $\text{CORPUS}_{\text{Aug}}$. This could be due to the high similarity between $\text{CORPUS}_{\text{Aug}}$ and $\text{CORPUS}_{\text{Ext}}$, as $\text{CORPUS}_{\text{Aug}}$ only contains a small fraction of documents from DUC datasets and both predominantly consist of Wikipedia documents. However, the fairly low precision and recall scores across all three collections underscore challenges in such large-scale knowledge retrieval. Addressing these challenges requires more research efforts to enhance retrieval performance.

7.2 Summarization Results

To answer **RQ2**, we make a comparison between our approach and baseline models, including weakly-supervised QFS, supervised RAG models, and LLM-based RAG models. The results are shown in Table 2. Interestingly, we find our

	CORPUS$_{Int}$					CORPUS$_{Ext}$					CORPUS$_{Aug}$				
Model	R1	R2	RS	BE	BA	R1	R2	RS	BE	BA	R1	R2	RS	BE	BA
Weakly Supervised															
QUERYSUM [51]	36.1	7.5	12.7	8.5	32.3	31.1	4.5	10.2	2.0	28.9	32.6	5.5	11.1	3.0	29.8
MARGESUM [52]	38.0	9.1	14.3	11.5	32.8	34.4	6.5	12.2	5.9	30.3	36.7	8.1	13.5	8.7	32.1
Supervised															
RAG-Sequence [32]	28.9	5.7	10.1	12.6	4.1	32.3	5.2	10.8	8.0	3.9	27.1	4.6	9.0	8.3	3.9
FiD [24]															
- BM25	42.4	11.3	16.5	21.4	38.1	38.8	8.4	14.2	15.7	35.0	41.4	10.8	16.1	20.0	36.8
- DPR	41.5	10.7	15.9	21.4	39.0	38.6	8.0	14.1	15.5	34.1	40.0	9.4	15.1	18.0	36.7
Zero-Shot Prompted															
NaiveRAG [11]															
- BM25	36.4	10.5	14.4	26.8	39.5	31.3	7.0	11.4	19.6	34.9	32.8	8.1	12.4	23.4	37.2
- DPR	37.1	10.4	14.7	27.3	39.5	31.7	7.1	11.4	18.3	34.8	33.7	8.4	12.6	21.8	37.1
Ours															
- BM25	43.1	11.0	16.5	25.9	40.7	37.9	7.3	13.1	19.9	34.5	39.4	8.8	14.3	22.1	37.2
- DPR	42.8	11.0	16.3	25.5	40.6	36.7	7.0	12.5	17.7	33.8	39.2	8.6	14.0	21.5	37.2
Few-Shot Prompted															
NaiveRAG [11]															
- BM25	41.7	11.9	16.5	24.4	41.7	35.2	7.8	12.8	17.1	35.5	37.5	9.1	14.2	20.2	37.7
- DPR	42.3	11.7	16.5	25.1	41.6	34.3	7.5	12.3	16.7	35.1	36.9	9.1	13.9	19.1	37.9
Ours															
- BM25	**45.8***	**13.4**	**18.6**	**29.2***	44.7	**41.7***	**9.6**	**15.6***	22.5	**37.3**	**43.6**	**11.3**	**16.7**	**26.1***	40.8
- DPR	45.1	12.9	18.3	28.6	**44.8***	41.1	9.4	15.2	**22.7***	36.8	42.9	10.7	16.2	24.7	**41.7***

Table 2: Summarization evaluation of our approach and baseline models over the three knowledge corpora. R1, R2, RS, BE, and BA stand for ROUGE-1, ROUGE-2, ROUGE-SU4, BERTScore, and BARTScore, respectively. The best results are in **bold**. *indicates that the improvement to the best baseline model is statistically significant (a two-paired t-test with p-value < 0.01).

few-shot prompted approach variant surpasses all baseline models across all evaluation metrics, especially in terms of semantic metrics. For instance, our best approach variant achieved a BERTScore of 29.2, surpassing the best-performing NaiveRAG, which had a BERTScore of 25.1. This result underscores the superiority of our proposed approach. Additionally, our approach consistently outperforms baselines across three different knowledge corpora and demonstrates robustness across various retrieval models. Furthermore, when comparing zero-shot and few-shot settings, we find that the performance of our approach in the few-shot setting significantly exceeds that in the zero-shot setting. This indicates that few-shot demonstrations have a highly positive effect on model performance. Lastly, our results reveal that fine-tuned FiD models outperform QFS models, indicating the advantages of fine-tuning on our training data. However, we also observe that the performance of the supervised RAG-Sequence model was comparatively lower, potentially due to the limited size of the training data.

7.3 Effect of Knowledge-Intensive Setup

To answer **RQ3**, we compare the performance of our best approach variant, which utilizes BM25, across original document sets (ORIGIN) and our three knowledge corpora. It is worth noting that the primary difference lies in the

Corpus	ROUGE-1	ROUGE-2	ROUGE-SU4	BERTScore	BARTScore
ORIGIN	49.8	17.3	21.9	32.5	47.6
$\text{CORPUS}_{\text{Int}}$	45.8	13.4	18.6	29.2	44.7
$\text{CORPUS}_{\text{Ext}}$	41.7	9.6	15.6	22.5	37.3
$\text{CORPUS}_{\text{Aug}}$	43.2	11.3	16.7	26.1	40.8

Table 3: Performance comparison of our approach variant with BM25 over the original document sets (ORIGIN) and our three knowledge corpora.

fact that ORIGIN contains only a limited set of manually curated relevant documents (around 40 documents) from original DUC datasets, whereas our knowledge corpora, such as $\text{CORPUS}_{\text{Ext}}$ and $\text{CORPUS}_{\text{Aug}}$, consists of millions of candidate documents. The results are shown in Table 3. We observe a significant decline in model performance when applied in the knowledge-intensive setup. For instance, when comparing ORIGIN and $\text{CORPUS}_{\text{Aug}}$, ROUGE-1 scores decrease by about 6 points, dropping from 49.8 to 43.2, and BERTScore scores decrease by about 6 points, dropping from 32.5 to 26.1. One possible explanation is that although both knowledge corpora offer substantial evidence for summarization, $\text{CORPUS}_{\text{Aug}}$ contains a considerable number of irrelevant documents. Consequently, retrieval models struggle to sift through this larger pool to identify the relatively small amount of relevant documents. This discrepancy underscores the significantly greater challenge posed by the knowledge-intensive setups compared to the traditional QFS, indicating a need for further research efforts to adapt retrieval models to these realistic scenarios.

8 Conclusion

This paper introduces a novel knowledge-intensive approach to QFS, aiming to address the limitations of traditional QFS setup that relies on the availability of relevant documents. This approach is tailored for practical scenarios involving highly specialized topics, eliminating the dependence on pre-existing document sets. The approach efficiently retrieves potentially relevant documents from a large-scale knowledge corpus based on the query. The summarization controller combines an LLM-based summarizer with a carefully tailored prompt, ensuring the quality of the generated summary. To assess the effectiveness of our proposed approach compared to strong baseline models, we introduce a specialized dataset, along with human-annotated relevance labels, to facilitate comprehensive retrieval and summarization evaluation. Extensive experiments demonstrate the superior performance of our method, particularly its ability to generate accurate summaries without relying on the availability of relevant documents initially. This underscores the versatility and practical applicability of our approach across diverse query scenarios, thereby contributing to advancements in QFS research.

References

1. Amplayo, R.K., Webster, K., et al.: Query refinement prompts for closed-book long-form QA. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7997–8012. Association for Computational Linguistics (2023)
2. Asai, A., Gardner, M., et al.: Evidentiality-guided generation for knowledge-intensive NLP tasks. In: NAACL-HLT. pp. 2226–2243 (2022)
3. Brown, T.B., Mann, B., et al.: Language Models are Few-Shot Learners. In: NeurIPS (2020)
4. Chen, D., Fisch, A., et al.: Reading Wikipedia to answer open-domain questions. In: ACL. pp. 1870–1879 (2017)
5. Craswell, N., Mitra, B., et al.: Overview of the TREC 2020 deep learning track. In: TREC. NIST Special Publication, vol. 1266 (2020)
6. Dang, H.T.: DUC 2005: Evaluation of question-focused summarization systems. In: Proceedings of the Workshop on Task-Focused Summarization and Question Answering. pp. 48–55 (2006)
7. Dinan, E., Roller, S., et al.: Wizard of wikipedia: Knowledge-powered conversational agents. In: ICLR (2019)
8. Fabbri, A., Li, I., et al.: Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In: ACL. pp. 1074–1084 (2019)
9. Fan, A., Jernite, Y., et al.: ELI5: Long form question answering. In: ACL. pp. 3558–3567 (2019)
10. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
11. Gao, Y., Xiong, Y., et al.: Retrieval-augmented generation for large language models: A survey. Arxiv (2023)
12. Giorgi, J., Soldaini, L., et al.: Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In: EMNLP Findings. pp. 8177–8199. Association for Computational Linguistics (2023)
13. He, P., Liu, X., et al.: Deberta: Decoding-Enhanced Bert with Disentangled Attention. In: ICLR (2021)
14. Huang, J.H., Alfadly, M., Ghanem, B., Worring, M.: Improving visual question answering models through robustness analysis and in-context learning with a chain of basic questions. arXiv preprint arXiv:2304.03147 (2023)
15. Huang, J.H., Dao, C.D., Alfadly, M., Ghanem, B.: A novel framework for robustness analysis of visual qa models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8449–8456 (2019)
16. Huang, J.H., Murn, L., Mrak, M., Worring, M.: Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In: Proceedings of the International Conference on Multimedia Retrieval. pp. 580–589 (2021)
17. Huang, J.H., Murn, L., Mrak, M., Worring, M.: Query-based video summarization with pseudo label supervision. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1430–1434. IEEE (2023)
18. Huang, J.H., Worring, M.: Query-controllable video summarization. In: Proceedings of the International Conference on Multimedia Retrieval. pp. 242–250 (2020)
19. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al.: Deepopht: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)

20. Huang, J.H., Yang, C.C., Shen, Y., Pacces, A.M., Kanoulas, E.: Optimizing numerical estimation and operational efficiency in the legal domain through large language models. In: ACM International Conference on Information and Knowledge Management (CIKM) (2024)
21. Huang, J.H., Yang, C.H.H., Chen, P.Y., Chen, M.H., Worring, M.: Causalainer: Causal explainer for automatic video summarization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2629–2635 (2023)
22. Huang, J.H., Zhu, H., Shen, Y., Rudinac, S., Pacces, A.M., Kanoulas, E.: A novel evaluation framework for image2text generation. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, LLM4Eval Workshop (2024)
23. Huang, K.H., Singh, S., et al.: SWING: Balancing coverage and faithfulness for dialogue summarization. In: EACL Findings. pp. 512–525 (2023)
24. Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open domain question answering. In: EACL. pp. 874–880 (2021)
25. Karpukhin, V., Oguz, B., et al.: Dense passage retrieval for open-domain question answering. In: EMNLP. pp. 6769–6781 (2020)
26. Kojima, T., Gu, S.S., et al.: Large language models are zero-shot reasoners. In: NeurIPS (2022)
27. Krishna, K., Roy, A., et al.: Hurdles to progress in long-form question answering. In: NAACL-HLT. pp. 4940–4957 (2021)
28. Kulkarni, S., Chammas, S., et al.: AQuaMuSe: Automatically generating datasets for query-based multi-document summarization. Arxiv **abs/2010.12694** (2020)
29. Laskar, M.T.R., Bari, M.S., et al.: A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In: ACL Findings. pp. 431–469 (2023)
30. Laskar, M.T.R., Hoque, E., et al.: WSL-DS: Weakly supervised learning with distant supervision for query focused multi-document abstractive summarization. In: COLING. pp. 5647–5654 (2020)
31. Lewis, M., Liu, Y., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL. pp. 7871–7880 (2020)
32. Lewis, P.S.H., Perez, E., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NeurIPS (2020)
33. Li, P., Wang, Z., et al.: Salience estimation via variational auto-encoders for multi-document summarization. In: AAAI. vol. 31 (2017)
34. Li, W., Zhang, X., et al.: Document-based question answering improves query-focused multi-document summarization. In: NLPCC. pp. 41–52 (2019)
35. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
36. Lin, J., Ma, X., et al.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: SIGIR. pp. 2356–2362 (2021)
37. Liu, Y., Wang, Z., et al.: QuerySum: A multi-document query-focused summarization dataset augmented with similar query clusters. In: AAAI. pp. 18725–18732 (2024)
38. Ma, C., Zhang, W.E., et al.: Multi-document Summarization via Deep Learning Techniques: A Survey. ACM Comput. Surv. **55**(5), 102:1–102:37 (2023)
39. Min, S., Michael, J., et al.: AmbigQA: Answering ambiguous open-domain questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5783–5797 (2020)

40. Nema, P., Khapra, M.M., et al.: Diversity driven attention model for query-based abstractive summarization. In: ACL. pp. 1063–1072 (2017)
41. Ouyang, L., Wu, J., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
42. Pasunuru, R., Celikyilmaz, A., et al.: Data augmentation for abstractive query-focused multi-document summarization. In: AAAI. pp. 13666–13674 (2021)
43. Petroni, F., Piktus, A., et al.: KILT: A benchmark for knowledge intensive language tasks. In: NAACL-HLT. pp. 2523–2544 (2021)
44. Raffel, C., Shazeer, N., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**, 140:1–140:67 (2020)
45. Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009)
46. Shaham, U., Segal, E., et al.: SCROLLS: Standardized CompaRison over long language sequences. In: EMNLP. pp. 12007–12021 (2022)
47. Stelmakh, I., Luan, Y., et al.: ASQA: Factoid questions meet long-form answers. In: EMNLP. pp. 8273–8288 (2022)
48. Thorne, J., Vlachos, A., et al.: FEVER: A large-scale dataset for fact extraction and VERification. In: NAACL-HLT. pp. 809–819 (2018)
49. Wang, A., Pang, R.Y., et al.: SQuALITY: Building a long-document summarization dataset the hard way. In: EMNLP. pp. 1139–1156 (2022)
50. Wei, J., Wang, X., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)
51. Xu, Y., Lapata, M.: Coarse-to-fine query focused multi-document summarization. In: EMNLP. pp. 3632–3645 (2020)
52. Xu, Y., Lapata, M.: Generating query focused summaries from query-free resources. In: ACL. pp. 6096–6109 (2021)
53. Yuan, W., Neubig, G., et al.: BARTScore: Evaluating Generated Text as Text Generation. In: NeurIPS. pp. 27263–27277 (2021)
54. Zhang, T., Kishore, V., et al.: BERTScore: Evaluating Text Generation with BERT. In: ICLR (2020)
55. Zhang, W., Aliannejadi, M., Yuan, Y., Pei, J., Huang, J.H., Kanoulas, E.: Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In: INLG (2024)
56. Zhang, W., Pal, V., Huang, J.H., Kanoulas, E., de Rijke, M.: QFMTS: Generating query-focused summaries over multi-table inputs. In: ECAI (2024)
57. Zhang, W., Vakulenko, S., Rajapakse, T., Kanoulas, E.: Scaling up query-focused summarization to meet open-domain question answering. ArXiv preprint, abs/2112.07536 (2021)
58. Zhang, W., Vakulenko, S., Rajapakse, T., Xu, Y., Kanoulas, E.: Tackling query-focused summarization as a knowledge-intensive task: A pilot study. The First Workshop on Generative Information Retrieval (Gen-IR) at SIGIR (2023)
59. Zhao, W.X., Liu, J., et al.: Dense text retrieval based on pretrained language models: A survey **42**, 1–60 (2024)
60. Zhao, Y., Qi, Z., et al.: QTSumm: Query-focused summarization over tabular data. In: EMNLP. pp. 1157–1172 (2023)
61. Zhong, M., Yin, D., et al.: QMSum: A new benchmark for query-based multi-domain meeting summarization. In: NAACL-HLT. pp. 5905–5921 (2021)
62. Zhou, Y., Shi, K., et al.: ODSum: New benchmarks for open domain multi-document summarization. Arxiv (2023)

63. Zhu, H., Huang, J.H., Rudinac, S., Kanoulas, E.: Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 978–987 (2024)

5. Neural Ranking with Weak Supervision for Open-Domain Question Answering: A Survey

Bibliographic Information

Xiaoyu Shen, **Svitlana Vakulenko**, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, Adrià de Gispert: Neural Ranking with Weak Supervision for Open-Domain Question Answering : A Survey. EACL (Findings) 2023: 1691-1705.

Neural Ranking with Weak Supervision for Open-Domain Question Answering : A Survey

Xiaoyu Shen[†1], **Svitlana Vakulenko**[1], **Marco del Tredici**[1], **Gianni Barlacchi**[1]
Bill Byrne[1,2] **and Adrià de Gispert**[1]
[1]Amazon Alexa AI
[2]Univeristy of Cambridge
†gyouu@amazon.com

Abstract

Neural ranking (NR) has become a key component for open-domain question-answering in order to access external knowledge. However, training a good NR model requires substantial amounts of relevance annotations, which is very costly to scale. To address this, a growing body of research works have been proposed to reduce the annotation cost by training the NR model with weak supervision (WS) instead. These works differ in what resources they require and employ a diverse set of WS signals to train the model. Understanding such differences is crucial for choosing the right WS technique. To facilitate this understanding, we provide a structured overview of standard WS signals used for training a NR model. Based on their required resources, we divide them into three main categories: (1) only documents are needed; (2) documents and questions are needed; and (3) documents and question-answer pairs are needed. For every WS signal, we review its general idea and choices. Promising directions are outlined for future research.

1 Introduction

Open-Domain Question Answering (ODQA) aims to provide precise answers in response to the user's questions by drawing on a large collection of documents (Voorhees et al., 1999). The majority of modern ODQA models follow the retrieve(-rerank)-read architecture: 1) given a question, a set of relevant documents are selected from a large document collection, and 2) the reader model produces an answer given this selected set and the question (Chen et al., 2017; Verga et al., 2021; Lee et al., 2021). Compared with parametric models without access to external knowledge, this architecture can better adapt to updated knowledge, offer easier interpretation and reduce hallucination (Zhu et al., 2021a; Shuster et al., 2021; Guo et al., 2022).

Conventional methods use sparse retrievers (SRs) such as TF-IDF and BM-25 in the first stage to match questions and documents via lexical overlap (Robertson and Walker, 1994), a process that may overlook semantically relevant documents with low lexical overlap with the question. Neural ranking (NR) models resolve this issue by encoding the questions and documents into dense vectors so that synonyms and paraphrases can be mapped to similar vectors through task-specific fine-tuning (Das et al., 2019; Karpukhin et al., 2020). However, training good NR models requires substantial amount of relevance annotations to perform competitively and NR models have been found to generalize poorly across domains (Thakur et al., 2021; Ren et al., 2022). In practice, collecting question-document relevance annotations is time-consuming. For a given question, an extensive annotation effort may be required to find the relevant documents. Repeating this annotation for every language and domain is not feasible (Shen et al., 2022c).

To reduce annotation costs, many techniques have been proposed to train NR models with *weak supervision (WS) signals* instead. This survey aims to provide a clear taxonomy to characterize these WS signals based on their required resources. There are three common resources that can be leveraged: (1) **Document** set, which is a bare minimum for building a ODQA system; (2) **Questions** without ground-truth relevance or answer annotations;

Resource	Weak-Supervision Signal
Documents (§3)	Self Contrastive Learning (§3.1)
	Question Generation (§3.2)
Documents +Questions (§4)	Sparse Retriever (§4)
	Pre-trained Language Model (§4)
	Supervised Teacher Model (§4)
Documents +QA Pairs (§5)	Answer as Document (§5.1)
	Answer-Document Mapping (§5.2)
	Latent-Variable Model (§5.3)

Table 1: Overview of different weak-supervision signals, together with their required resources, that we can apply to train NR models for open-domain question-answering.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1736–1750

(3) **Question-Answer (QA) Pairs**, which are a set of already answered questions. Different applications have different levels of resource availability. For example, common domains such as e-commerce normally already have large amounts of question-answer pairs from customer services while smaller domains or low-resource languages can only have a document set without any existing questions. For each level of resource availability, we review the applicable WS signals. An overview can be seen in Table 1.

While there have been surveys that describe general neural information retrieval (IR) approaches (Mitra et al., 2018; Guo et al., 2020; Lin et al., 2021a; Zhu et al., 2021a; Guo et al., 2022), we focus specifically on the low-resource scenarios, which makes our contribution unique in this respect. The closest to our work is the BEIR benchmark for zero-shot cross-domain evaluation of IR models (Thakur et al., 2021) and its multiple related studies (Mokrii et al., 2021; Reddy et al., 2021; Wang et al., 2022; Ren et al., 2022). Nonetheless, these studies test specific algorithms but do not provide a holistic overview of how they are related. Our survey can be useful in that: (1) future WS research can use it as a reference book to compare with similar techniques. (2) It can serve as a practical guide for choosing the best WS sginals to train NR models given different availability levels of resources. (3) As the retrieve(-rerank)-read paradigm is generic and has been increasingly popular across NLP tasks like machine translation (He et al., 2021) and intent detection (Mehri and Eric, 2021), it can have broader impact in many other applications. Therefore, although this survey illustrates with the use case of ODQA, *the introduced techniques are intended to go beyond specific applications*.

In the following sections we first lay out the necessary background knowledge (§2), then explain popular WS signals when different resources are available in sections 3 to 5. In conclusion, we highlight promising directions for future work (§6).

2 Background

Neural Ranking (NR) for ODQA Let Q, D and A denote the question, document and answer set. Given a question $q \in Q$, the NR model assigns a relevance score $\mathcal{R}(q, d)$ to each $d \in D$ and selects top-k document $D_{topk} \in D$ with the highest relevance scores. Afterwards, a reader will estimate the score $\mathcal{G}(a|q, D_{topk})$ to predict the final answer $a \in A$ conditioned on both q and D_{topk}. The NR model can be implemented using various architecture with increasing model complexity. For computational efficiency, normally a bi-encoder architecture (Bromley et al., 1993) is first applied to pre-select top candidates from the whole document set, then a more complex cross-interaction model is applied to provide more accurate relevance scores only for the preselected candidates (Lee et al., 2021). The training objective for the NR model $\mathcal{R}$ can be formalized as:

$$\min_{\mathcal{R}} \mathbb{E}_{q,d^+,d^-_{1\sim n} \in Q\times D} \mathcal{L}(\mathcal{R}, q, d^+, d^-_{1\sim n}) \quad (1)$$

where $Q \times D$ indicates the full set of question-document pairs, d^+ is a positive (relevant) document for q, $d^-_{1\sim n}$ is the sampled n negative (irrelevant) documents and $\mathcal{L}$ is the loss function. A common choice for $\mathcal{L}$ is the contrastive loss:

$$\mathcal{L} = -\log \frac{e^{\mathcal{R}(q,d^+)}}{e^{\mathcal{R}(q,d^+)} + \sum_{j=1}^{n} e^{\mathcal{R}(q,d^-_j)}} \quad (2)$$

Neural Ranking with Weak Supervision In the standard supervised setting we need relevance annotations for $(q, d) \rightarrow \{+, -\}$ to train $\mathcal{R}$ with Eq 1. Obtaining high-quality relevance annotations requires tremendous human labor and is expensive to scale to multiple domains (Del Tredici et al., 2021; Ram et al., 2022). Weak supervision (WS) is a widely-used approach to reduce such cost by leveraging supervision signals from e,g., heuristic rules, knowledge bases or external models (Zhang et al., 2021). WS signals are cheap to obtain but might contain significant noise which will affect the NR performance. Therefore, understanding their working mechanisms and pros and cons are important to obtain a good NR model. We group WS signals into 3 classes by the resources that they need: (1) *Documents*: only document collection D is needed; (2) *Documents + Questions*: document collection D and question set Q are needed; (3) *Documents + QA Pairs*: document collection D and QA pairs (Q, A) are needed. In the next section, we will present the three classes of WS signals and discuss their pros and cons.

3 Resource: Documents

This section discusses two main techniques to produce WS signals requiring only the document set: (1) self-contrastive learning and (2) question generation. This makes the minimum assumption about

Method	Pseudo Question
Perturbation	d with added perturbation
Summary	(Pseudo) summary of d
Proximity	Nearby text of d
Cooccurrence	Text sharing cooccurred spans with d
Hyperlink	Text with a hyperlink to/from d

Table 2: Given a document d, different heuristics to construct pseudo questions. Contrastive samples made from these heuristics serve as WS signals to train the NR model.

resource availability since having the document set is a prerequisite for building an ODQA system.

3.1 Self Contrastive Learning

Self contrastive learning relies on heuristics to construct pseudo question-document pairs $(q', d'^{+/-})$ from D, then uses them to supervise training of a NR model. The objective is:

$$\min_{\mathcal{R}} \mathbb{E}_{q', d'^{+}, d'^{-}_{1\sim n} \in D} \mathcal{L}(\mathcal{R}, q', d'^{+}, d'^{-}_{1\sim n}) \quad (3)$$

where $\mathcal{L}$ is the ranking loss as in Eq 1. Since negative pairs can be easily constructed by random sampling, the main difficulty is to design good heuristics for constructing positive pseudo pairs (q', d'^{+}). There are 5 popular heuristics to construct such positive pairs: perturbation-based, summary-based, proximity-based, cooccurence-based and hyperlink-based. An overview is in Table 2.

Perturbation-based heuristics add perturbations to some text, then treat the perturbed text and the original text as a positive pair. The intuition is that *perturbed text should still be relevant to the original text*. Typical choices of perturbations include word deletion, substitution and permutation (Zhu et al., 2021b; Meng et al., 2021), adding drop out to representation layers (Gao et al., 2021), or passing sentences through different language models (Carlsson et al., 2021), among other.

Summary-based heuristics extract a summary from the document as the pseudo question based on the intuition that *questions should contain representative information about the central topic of the document*. The summary can be the document title (MacAvaney et al., 2017, 2019; Mass and Roitman, 2020), a random sentence from the first section of the document (Chang et al., 2020), randomly sampled ngrams (Gysel et al., 2018) or a set of keywords generated from a document language model (Ma et al., 2021a).

Proximity-based heuristics utilize the position information in the document to obtain positive pairs based on the intuition that *nearby text should be more relevant to each other*. The most famous one is the inverse-cloze task (Lee et al., 2019), where a sentence from a passage is treated as the question and the original passage, after removing the sentence, is treated as a positive document. They can be combined with typical noise injection methods like adding drop-out masks (Xu et al., 2022), random word chopping or deletion (Izacard et al., 2021) to further improve the model robustness. Other methods include using spans from the same document (Gao and Callan, 2022; Ma et al., 2022), sentences from the same paragraph, paragraphs from the same document as positive samples (Di Liello et al., 2022), etc.

Cooccurrence-based heuristics construct positive samples based on the intuition that *sentences containing cooccurred spans are more likely to be relevant* (Ram et al., 2021). For example, Glass et al. (2020) constructs a pseudo question with a sentence from the corpus. A term from it is treated as the answer and replaced with a special token. Passages retrieved with BM25 which also contains the answer term are treated as pseudo positive documents. Ram et al. (2022) treat a span and its surrounding context as the pseudo question and use another passage that contains the same span as a positive document.

Hyperlink-based heuristics leverage hyperlink information based on the intuition that *hyperlinked text are more likely to be relevant* (Zhang et al., 2020; Ma et al., 2021b). For example, Chang et al. (2020) takes a sentence from the first section of a page p as a pseudo question because it is often the description or summary of the topic. A passage from another page containing hyperlinks to p is treated as a positive document. Yue et al. (2022a) replace an entity word with a question phrase like "what/when" to form a pseudo question. A passage from its hyperlinked document that contains the same entity word is treated as a positive sample. Zhou et al. (2022) build positive samples with two typologies: "dual-link" where two passages have hyperlinks pointed to each other, and "co-mention" where two passages both have a hyperlink to the same third-party document.

Input	Document Document + Answer Document + Answer + Question Type Document + Answer + Question Type + Clue
Question Generator	Rule-based Generator Prompt-based Generator Fine-tuned Generator$^{\diamond}$
Filter	LM Score Round-trip Consistency Probability from pre-trained QA Influence Function Ensemble Consistency Entailment Score Learning to Reweight$^{\diamond}$ Target-domain Value Estimation$^{\diamond}$

Table 3: Different choices for a question generation model setup. $^{\diamond}$ means minimal relevance annotations are needed.

3.2 Question Generation

Self contrastive learning relies on sentences already present in D. Question generation leverage a question generator (QG) to generate new questions *not found* in D, which can then be used to provide WS signals for the NR model. It often employs a filter Fil to filter poorly generated questions. The training objective is:

$$\min_{\mathcal{R}} \mathbb{E}_{q=QG(d^+)\&Fil(q,d^+)=0} \mathcal{L}(\mathcal{R}, q, d^+, d^-_{1\sim n})$$

where the expectation is with respect to documents d^+ and $d^-_{1\sim n}$ drawn from D, the q are generated from d^+, $Fil(q, d^+) = 0$ requires that these questions not be discarded by Fil, and $\mathcal{L}$ is the standard ranking loss. There are various ways of designing the question generator and filter. We will cover the popular choices in the following section. An overview can be seen in Table 3.

Choices of Input A variety of information can be provided as input for the QG. The most straightforward approach is answer-agnostic which provides only the document (Du and Cardie, 2017; Kumar et al., 2019). In this way, the model can choose to attend to different spans of the document as potential answers and so generate different, corresponding questions. A more common method is answer-aware where an answer span is first extracted from a document, then the QG generates a question based on both the document and answer (Alberti et al., 2019; Shakeri et al., 2020). Finer-grained information can also be provided such as the question type ("what/how/...") (Cao and Wang, 2021; Gao et al., 2022) as well as additional clues (such as document context to disambiguate the question) (Liu et al., 2020). Adding more information reduces the entropy of the question and makes it easier for the model to learn, but also increases the possibility of error propagation (Zhang and Bansal, 2019). In practice, well-defined filters should be applied to remove low-quality questions.

Choices of Question Generator There are three popular choices for the question generator. (1) Rule-based methods (Pandey and Rajeswari, 2013; Rakangor and Ghodasara, 2015) rely on handcrafted templates and features. These are time-consuming to design, domain-specific, and can only cover certain forms of questions. (2) Prompt-based methods relying on pre-trained language models (PLMs). Documents can be presented to a PLM, with an appended prompt such as "*Please write a question based on this passage*" so that the PLM can continue the generation to produce a question (Bonifacio et al., 2022; Sachan et al., 2022; Dai et al., 2022). (3) Fine-tuned generators that are trained on annotated question-document pairs. When in-domain annotations are not enough, we can leverage out-of-domain (OOD) annotations, if any, to fine-tune the QG. The first two QGs require no training data, but their quality is often inadequate. In practice, we should only consider them when *there is a complete lack of high-quality supervised data for fine-tuning the QG*. When target-domain questions are available, we can also apply semi-supervised techniques such as back-training to adapt the QG to the target domain (Zhao et al., 2019; Kulshreshtha et al., 2021; Shen et al., 2022a).

Choices of Filter Filtering is a crucial part of QG since a significant portion of generated questions could be of low quality and would provide misleading signals when used to train the NR model (Alberti et al., 2019). A typical choice is filtering based on round-trip consistency (Alberti et al., 2019; Dong et al., 2019), where a pre-trained QA system is applied to produce an answer based on the generated question. A question is kept only when the produced answer is consistent with the answer from which the question is generated. We can also relax this strict consistency requirement and manually adjust an acceptance threshold based on the probability from the pre-trained QA system (Zhang and Bansal, 2019; Lewis et al., 2021), LM score from the generator itself (Shakeri et al., 2020; Liang et al., 2020), or an entailment score from a model trained on question-context-answer pairs (Liu et al., 2020). Influence functions (Cook and Weisberg, 1982) can be used to estimate the

effect on the validation loss of including a synthetic example (Yang et al., 2020), but this does not achieve satisfying performances on QA tasks (Bartolo et al., 2021). Bartolo et al. (2021) propose filtering questions based on ensemble consistency, where an ensemble of QA models are trained with different random seeds and only questions agreed by most QA models are selected. When minimal target-domain annotation is available, we can also learn to reweight pseudo samples based on the validation loss (Sun et al., 2021), or use RL to select samples that lead to validation performance gains (value estimation) (Yue et al., 2022b).

3.3 Discussion

If the heuristics or QG are properly designed, NR models trained from their supervision can even match the fully-supervised performance (Wang et al., 2022; Ren et al., 2022). The biggest challenge is the difficulty to pick the most suitable heuristics or QG when we face a new domain. A general solution is to automatically select good pseudo pairs with reinforcement learning (RL) when minimal target-domain annotations are available (Zhang et al., 2020), so as avoiding the need to manually fixing the WS signals, but this would bring significant computational overhead. In practice hyperlink-based approaches often perform the best among the heuristics as they have additional reference information to leverage, which makes them most similar to the actual relevance annotations. However, hyperlink information is not available in most domains and thereby limits its use cases (Sun et al., 2021). QG-based WS signals are often preferred over heuristics-based ones as they can produce naturally-sound questions themselves without relying on the chance to find good pseudo questions in the documents. Nonetheless, obtaining a high-performing QG can also be non-trivial. One big challenge comes from the one-to-many mapping relations between questions and documents. Under this situation, standard supervised learning tends to produce safe questions with less diversity and high lexical overlap with the document. For example, Shinoda et al. (2021) found that QG reinforces the model bias towards high lexical overlap. We will need more sophisticated training techniques such as latent-variable models (Shen and Su, 2018; Xu et al., 2020; Li et al., 2022) and reinforcement learning (Yuan et al., 2017; Zhang and Bansal, 2019; Shen et al., 2019a) to alleviate the model bias towards safe questions.

4 Resource: Documents + Questions

This section includes WS signals that require additional access to a question set Q. In practice, annotating question-document relations usually requires domain experts to read long documents and careful sampling strategies to ensure enough positive samples, while unlabeled questions are much easier to obtain either through real user-generated content or simulated traffic. Therefore, it is common to have a predominance of unlabeled questions. The crucial point is to establish the missing relevance labels. Suppose a WS method can provide the missing label $\text{WS}(q, d)$ for a question-document pair (q, d), then we can use it to supervise the NR model by:

$$\min_{\mathcal{R}} \mathbb{E}_{q \in Q, d \in D} \mathcal{L}(\mathcal{R}(q, d), \text{WS}(q, d)) \quad (4)$$

where $\mathcal{L}$ is the loss function that encourages similarity between $\mathcal{R}(q, d)$ and $\text{WS}(q, d)$.

There are three popular types models that can provide such WS signal here: (1) sparse retriever, (2) pre-trained language model and (3) supervised teacher model.

Sparse Retriever (SR) Recent research finds that NR and SR models are complementary. NR models are better at semantic matching while SRs are better at capturing exact match and handling long documents (Chen et al., 2021; Luan et al., 2021). SRs are also more robust across domains (Thakur et al., 2021; Chen et al., 2022). This motivates the use of unsupervised sparse retrievers like BM25 as WS signals. For example, Dehghani et al. (2017); Nie et al. (2018) train a NR model on samples annotated with BM25. Xu et al. (2019) apply four scoring functions to auto-label questions and documents with: (1) BM25 scores, (2) TF-IDF scores, (3) cosine similarity of universal embedding representation (Cer et al., 2018) and (4) cosine similarity of the last hidden layer activation of pre-trained BERT model (Devlin et al., 2019). Both papers observe that the resulting model outperforms BM25 on the test sets. Chen et al. (2021) further show that distilling knowledge from BM25 helps the retriever to better match rare entities and improves zero-shot out-of-domain performance.

Pre-trained Language Model (PLM) As PLMs already encode significant linguistic knowledge, there have also been attempts at using prompt-based PLMs to provide WS signals for question-

document relations (Smith et al., 2022; Zeng et al., 2022). Similar as in question generation, we can use prompts like "*Please write a question based on this passage*", concatenate the document and question, then use the probability assigned by the PLM to auto-label question-document pairs. To maximize the chances of finding positive document, normally we first obtain a set of candidate documents by BM25, then apply PLM to auto-label the candidate set (Sachan et al., 2022). This can further exploit the latent knowledge inside PLMs that has been honed through pre-training, so it often shows better performance compared with weak supervision only using BM25 (Nogueira et al., 2020; Singh Sachan et al., 2022).

Supervised Teacher Model A very common choice is using a supervised teacher model to provide WS signals. The teacher model is "supervised" because it is explicitly fine-tuned on annotated question-document pairs. When in-domain annotations are not sufficient, we can leverage out-of-domain (OOD) annotations, if available, to train the teacher model. The teacher model usually employs a more powerful architecture such as with more complex interactions or larger sizes. It may not be directly applicable in downstream tasks due to the latency constraints, but can be useful in providing WS signals for training the NR model. For example, previous research has shown that models with larger sizes or late/cross-interaction structures generalize much better on OOD data (Pradeep et al., 2020; Lu et al., 2021; Ni et al., 2021; Rosa et al., 2022; Muennighoff, 2022; Zhan et al., 2022). After training a teacher model on OOD annotations, applying it to provide WS signals through target-domain question and document collections can significantly improve the in-domain performance of the NR model (Hofstätter et al., 2021; Lin et al., 2021b; Lu et al., 2022). Kim et al. (2022) further show that we can even use the same architecture and capacity to obtain a good teacher model. They expand the question with centroids of word embeddings from top retrieved passages (using BM25), and then use the expanded query for self knowledge distillation. Similar ideas of reusing the same architecture to provide WS signals have also been explored by Yu et al. (2021a); Kulshreshtha et al. (2021); Zhuang and Zuccon (2022).

Discussion The three WS signals listed above work directly on actual questions instead of pseudo pairs as in §3 so that the NR model can adapt better to the target-domain question distribution. The bottleneck is the quality of the WS signals. SRs and PLMs are unsupervised, which could be more robust when we face a completely different domain (Dai et al., 2022). Otherwise, if we already have certain amounts relevance annotations from the target or similar domains, usually using a supervised teacher model is preferred. Nevertheless, these WS signals inevitably contain noise, and can harm the downstream performance if the noise is significant. There are two main strategies to reduce the noise effects: (1) Apply less strict margin-based loss such as the hinge loss (Dehghani et al., 2017; Xu et al., 2019) and MarginMSE loss (Hofstätter et al., 2020; Wang et al., 2022), then models have fewer chances of overfitting to the exact labels, and (2) Apply noise-resistant training methods such as confidence-based filtering (Mukherjee and Awadallah, 2020; Yu et al., 2021b) and meta-learning-based refinement (Ren et al., 2018; Zhu et al., 2022). Another potential issue is that the amount of training data in this section relies on the amount of questions we have. Unlike the document set which we can obtain for free, the question set takes time to collect and are often orders of magnitudes smaller. If no sufficient questions are available, we can use synthetic questions from question generation, then apply same WS signals in this section, which has been shown to perform on par with using real questions in certain domains (Wang et al., 2020, 2022; Thakur et al., 2022).

5 Resource: Documents + QA Pairs

Many domains have large numbers of already answered questions from customer services, technical support or web forums (Huber et al., 2021). These QA pairs can provide richer information than only unlabeled questions. However, most answers are based on personal knowledge, derived from experience, and do not include a reference to any external document. This prevents their direct use as training data for the NR model. This section introduces three standard methods that exploit QA pairs to provide WS signals despite this difficulty: (1) Answer as document, (2) Answer-document mapping and (3) Latent-variable models.

5.1 Answer as a Document

As a straightforward way to leverage QA pairs, this method directly treats QA pairs as positive samples

and does not distinguish between documents and answers (Lai et al., 2018). These QA pairs can provide direct WS signals to train the NR model:

$$\min_{\mathcal{R}} \mathbb{E}_{q,a^+,a^-_{1\sim n}\in Q\times A}\mathcal{L}(\mathcal{R}, q, a^+, a^-_{1\sim n}) \quad (5)$$

where $(q, a^+) \in Q \times A$ are question-answer pairs in the target domain, $a^-_{1\sim n}$ are sampled n negative answers and $\mathcal{L}$ is the standard ranking loss.

Though simple, this has been a common practice to "warm up" the NR model when no sufficient relevance annotations are available. For large-sized models, this can be crucial to fully leverage the model capacity since we often have orders of magnitude more QA pairs than relevance annotations (Ni et al., 2021; Oğuz et al., 2021). However, the style, structure and format differ between the document and the answer. The answer is a direct response to the question, and so it is easier to predict due to its strong semantic correlation with the question. Whereas the document can be implicit and may contain fewer obvious clues that can imply an answer; deep text understanding is required to predict the relevance between questions and documents (Zhao et al., 2021; Shen et al., 2022b). Therefore, this approach may be insufficient to reach satisfying results as a standalone method.

5.2 Answer-document Mapping

This approach leverages an additional mapping function to automatically link answers to the corresponding documents. The NR model can get WS signals from the linked answers:

$$\min_{\mathcal{R}} \mathbb{E}_{q,a\in(Q,A),d^-_{1\sim n}\sim D}\mathcal{L}(\mathcal{R}, q, M(a), d^-_{1\sim n})$$

where $(q, a) \in (Q, A)$ are question-answer pairs, M is a mapping function from an answer to its corresponding document, and $\mathcal{L}$ is the standard ranking loss. The mapping function is based on hand-crafted heuristics. For long-form descriptive answers, a popular way is to map them to documents with highest ROUGE scores (Lin, 2004) since the answers can be considered as summaries of the original documents (Fan et al., 2019). For short-span answers, a popular way is to map them to top-ranked documents retrieved using BM25 that contain the answer span (Karpukhin et al., 2020; Sachan et al., 2021; Christmann et al., 2022).

Answer-document mapping was widely adopted for constructing large-scale datasets in information retrieval (Joshi et al., 2017; Dunn et al., 2017; Elgohary et al., 2018). This can work well if the mapping has high accuracy, which is often difficult to achieve. Frequent answers or entities might lead to false positive mappings. It is also difficult to find positive documents for boolean and abstractive answers using only heuristics-based mapping functions (Izacard and Grave, 2021). Models can easily overfit to the biases introduced via such mapping function (Du et al., 2022).

Distribution of $\mathcal{R}(z\|q)$	**Optimization Method**
Categorical	Top-k approximation
Multinomial	EM algorithm Learning from attention

Table 4: Distribution assumptions made about the neural ranker and corresponding optimization methods, suppose we train the NR model following Equation 6.

5.3 Latent-Variable Model

We can still train the NR model on question-document pairs as in Answer-Document Mapping. However, instead of relying on a heuristic-based mapping function, we can treat this mapping as a "latent variable" within a probabilistic generative process (Lee et al., 2019; Shen, 2022). By this means, the NR model $\mathcal{R}$ gets WS signals from the QA reader $\mathcal{G}$ by maximizing the marginal likelihood:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)} \log \sum_{z\sim Z} \mathcal{R}(z|q)\mathcal{G}(a|q,z) \quad (6)$$

where Z indicates all possible document combinations. Directly optimizing over Eq 6 is infeasible as it requires enumerating over all documents. A closed-form solution does not exist due to the deep neural network parameterization of $\mathcal{R}$ and $\mathcal{G}$. The following section explains popular optimization options. An overview can be seen in Table 4.

Top-k approximation A popular approach is to assume a categorical distribution for $\mathcal{R}(Z|q)$; that is, to assume for each question only a single document is selected and the answer is generated from that one document. Eq 6 can be approximated by enumerating over only the top-k documents, assuming the remaining documents having negligibly small contributions to the likelihood:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{D}_{q,a\in(Q,A)} \log \sum_{z\sim E_{topk}} \mathcal{R}(z|q)\mathcal{G}(a|q,z)$$

This has been a popular choice in end-to-end training of text generation models (Lee et al., 2019; Shen et al., 2019b; Guu et al., 2020; Lewis et al.,

2020; Shuster et al., 2021; Ferguson et al., 2022). Despite its simplicity, the top-k approximation has two main drawbacks. (1) The approximation is performed on the top-k documents obtained from the NR model. If the NR model is very weak at the beginning of training, these top-k documents can be a bad approximation to the real joint likelihood and the model might struggle to converge. (2) The assumption that document follow a categorical distribution might be problematic especially if the answer requires evidence from multiple documents (Wang and Pan, 2022).

Expectation–Maximization (EM) algorithm To address the second drawback of the top-k approximation approach, we can assume a multinomial distribution for $R(Z|q)$ so that an answer can be generated from multiple documents. The cost of this relaxation is the increased difficulty of optimization. Approximating the joint likelihood from top-k samples becomes infeasible due to the combinatorial distribution of document. Singh et al. (2021) propose optimizating it with the EM algorithm under an independent assumption about the posterior distribution of $\mathcal{R}(z|q)$:

$$\max_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)}[\log \sum_{z\in D_{topk}} \mathcal{R}(z|q) \times SG(\mathcal{G}(a|q,z)) + \log \mathcal{G}(a|q,D_{topk})] \quad (7)$$

where SG means stop-gradient (gradients are not backpropagated through $\mathcal{G}$). As can be seen, the training signal for the NR model is essentially the same as in the *Top-k Approximation* case, except that the reader is trained by conditioning on all top-k documents to generate the answer. Singh et al. (2021) also find that Eq 7 is quite robust with respect to parameter initialization. Similarly, Zhao et al. (2021) apply the hard-EM algorithm to train the NR model, which only treats documents with the highest likelihood estimated by the reader as positive. Izacard et al. (2022) further experiment with using the leave-one-out perplexity from the reader to supervise the ranker.

Learning from attention Another way to optimize the NR model in Eq 6 is to leverage attention scores from the reader $\mathcal{G}$. The assumption is that when training $\mathcal{G}$ to generate the answer, its attention score is a good approximation of question-document relevance. The training objective is:

$$\min_{\mathcal{R},\mathcal{G}} \mathbb{E}_{q,a\in(Q,A)} \sum_{z\sim E_{topk}} \mathcal{L}(A_z|\mathcal{R}(z|q)) - \log \mathcal{G}(a|q,Z=D_{topk}) \quad (8)$$

where $\mathcal{G}$ is trained to generate the right answer based on the question and the top-k document, same as in the EM algorithm. A_z is the attention score of $\mathcal{G}$ on the document z. $\mathcal{L}$ is the loss function to encourage the similarity between distributions of the attention scores and retrieving scores.

Izacard and Grave (2021) propose a training process that optimizes $\mathcal{R}$ and $\mathcal{G}$ iteratively. $\mathcal{R}$ is trained to minimize KL divergence between relevance and attention scores. (Lee et al., 2021) jointly optimize $\mathcal{R}$ and $\mathcal{G}$ and apply a stop-gradient operation on $\mathcal{G}$ when updating $\mathcal{R}$. Sachan et al. (2021) use retriever scores to bias attention scores on the contrary. These can be considered as first-order Taylor series approximations of Eq. 6 by replacing $\mathcal{R}(Z|q)$ with attention scores (Deng et al., 2018).

Discussion Training with latent-variable models can perform close to fully supervised models under certain scenarios (Zhao et al., 2021; Sachan et al., 2021). The main challenge is the training difficulty. In practice, we can often initialize the NR model using the *answer as document* or *answer-document mapping* to make the training more stable. If not enough QA pairs are available, we can use heuristics like masked salient entities (Guu et al., 2020) to form pseudo pairs, then apply the same WS techniques in this section. Combining supervision signals from various various optimization techniques such as learning from attention and EM algorithm can also be beneficial (Izacard et al., 2022). If the independence assumption made by Eq 7 does not hold, we need to resort to more complex optimization algorithms. A potential direction is to apply a Dirichlet prior over $R(z|q_t)$, which is a conjugate distribution to the multinomial distribution (Minka, 2000), with the result that the sampled document are not independent individuals but a combination set. Eq 6 can then be estimated by rejection sampling (Deng et al., 2018) or a Laplace approximation (Srivastava and Sutton, 2017) so as to avoid the independence assumption about the posterior distribution. Nonetheless, this will further increase the training complexity, which is already a key bottleneck for training the NR model.

6 Conclusions

We review standard WS signals used for training NR models in ODQA and provide a structured way of classifying them according to the required resource. For WS signal, we discuss different options and summarize the pros and cons. As a final wrap-up, we list promising directions that we believe worth exploring further: (1) *How to select the most suitable technique for a given scenario?* Despite the wide range of applicable techniques, it is non-trivial to decide how to select the best one except for an empirical experimentation. (2) *To which extent are these techniques complementary?* Existing work compares performance only between similar types of methods but not across the whole range of techniques and resources available. This makes it hard to decide whether different approaches could potentially complement each other and how they should be combined effectively. (3) *Do methods work across languages?* The vast majority of current research is conducted on English datasets. Even though all described methods in this survey have no explicit restrictions on languages they can be applied to, it is likely that their performance will vary across languages, especially for the methods relying on handcrafted heuristics.

Limitations

This survey covers introductions and related work of major WS algorithms used for neural ranking. Due to the space limit, most methods included in this paper are brief. Readers might not have a good understand on all the introduced methods. Interested readers can refer to existing surveys about general knowledge in QA (Zeng et al., 2020; Zhu et al., 2021a; Roy and Anand, 2021; Rogers et al., 2021; Pandya and Bhatt, 2021). Furthermore, we did not provide points to existing ODQA datasets and the performance of recent models. The conclusions in this survey also come from summaries of previous works. The lack of datasets including various resources needed for different WS algorithms prevents a comprehensive, fair comparison across algorithms. We hope future research can work on the creation of more datasets with various availabilitis of resources in different domains to enable this comparison. Lastly, we aim to create a big picture from the technology level, so we did not strictly limit our references only to the application of ODQA. The connection to specific ODQA applications might be loose, readers would need to extract useful information for the specific use cases.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. *arXiv preprint arXiv:2201.10582*.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? *arXiv preprint arXiv:2110.06918*.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational question answering on heterogeneous sources. *arXiv preprint arXiv:2204.11677*.

R Dennis Cook and Sanford Weisberg. 1982. *Residuals and influence in regression*. New York: Chapman and Hall.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adriá de Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2974–2978.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. *Advances in Neural Information Processing Systems*, 31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Pre-training transformer models with sentence-level objectives for answer sentence selection. *arXiv preprint arXiv:2205.10455*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, 32.

Pan Du, Jian-Yun Nie Nie, Yutao Zhu, Hao Jiang, Lixin Zou, and Xiaohui Yan. 2022. Pregan: Answer oriented passage ranking with weakly supervised gan. *arXiv preprint arXiv:2207.01762*.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1083.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

James Ferguson, Hannaneh Hajishirzi, Pradeep Dasigi, and Tushar Khot. 2022. Retrieval data augmentation informed by downstream question answering performance. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 1–5, Dublin, Ireland. Association for Computational Linguistics.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2022. " what makes a question inquisitive?" a study on type-controlled inquisitive question generation. *arXiv preprint arXiv:2205.08056*.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, GP Shrivatsa Bhargav, Dinesh Garg, and Avirup Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Christophe Van Gysel, Maarten De Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Transactions on Information Systems (TOIS)*, 36(4):1–25.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Patrick Huber, Armen Aghajanyan, Barlas Oğuz, Dmytro Okhonko, Wen-tau Yih, Sonal Gupta, and Xilun Chen. 2021. Ccqa: A new web-scale question answering dataset for model pre-training. *arXiv preprint arXiv:2110.07731*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. *ICLR*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jihyuk Kim, Minsoo Kim, and Seung-won Hwang. 2022. Collective relevance labeling for passage retrieval. *arXiv preprint arXiv:2205.03273*.

Devang Kulshreshtha, Robert Belfer, Iulian Vlad Serban, and Siva Reddy. 2021. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7064–7078.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872.

Tuan Manh Lai, Trung Bui, Nedim Lipka, and Sheng Li. 2018. Supervised transfer learning for product information question answering. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1109–1114. IEEE.

Haejun Lee, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher D Manning, and Kyoung-Gu Woo. 2021. You only need one model for open-domain question answering. *arXiv preprint arXiv:2112.07381*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Jin Li, Peng Qi, and Hong Luo. 2022. Generating consistent and diverse qa pairs from contexts with bn conditional vae. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 944–949. IEEE.

Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021a. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, pages 2032–2043.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. *arXiv preprint arXiv:2205.09153*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Pre-train a discriminative text encoder for dense retrieval via contrastive span prediction. *arXiv preprint arXiv:2204.10641*.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021a. Prop: Pre-training with representative words prediction for ad-hoc retrieval. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 283–291.

Zhengyi Ma, Zhicheng Dou, Wei Xu, Xinyu Zhang, Hao Jiang, Zhao Cao, and Ji-Rong Wen. 2021b. Pre-training for ad-hoc retrieval: Hyperlink is also you need. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1212–1221.

Sean MacAvaney, Kai Hui, and Andrew Yates. 2017. An approach for weakly-supervised deep information retrieval. *arXiv preprint arXiv:1707.00189*.

Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019. Content-based weak supervision for ad-hoc re-ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–996.

Yosi Mass and Haggai Roitman. 2020. Ad-hoc document retrieval using weak-supervision with bert and gpt2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4191–4197.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Thomas Minka. 2000. Estimating a dirichlet distribution.

Bhaskar Mitra, Nick Craswell, et al. 2018. *An introduction to neural information retrieval*. Now Foundations and Trends Boston, MA.

Iurii Mokrii, Leonid Boytsov, and Pavel Braslavski. 2021. A systematic evaluation of transfer learning and pseudo-labeling with bert-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2081–2085.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems*, 33:21199–21212.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Yifan Nie, Alessandro Sordoni, and Jian-Yun Nie. 2018. Multi-level abstraction convolutional model with weak supervision for information retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 985–988.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Barlas Oğuz, Kushal Lakhotia, Anchit Gupta, Patrick Lewis, Vladimir Karpukhin, Aleksandra Piktus, Xilun Chen, Sebastian Riedel, Wen-tau Yih, Sonal Gupta, et al. 2021. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*.

Shivank Pandey and KC Rajeswari. 2013. Automatic question generation using software agents for technical institutions. *International Journal of Advanced Computer Research*, 3(4):307.

Hariom A Pandya and Brijesh S Bhatt. 2021. Question answering survey: Directions, challenges, datasets, evaluation matrices. *arXiv preprint arXiv:2112.03572*.

Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. 2020. H2oloo at trec 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. *Corpus*, 5(d3):d2.

Sheetal Rakangor and YR Ghodasara. 2015. Literature review of automatic question generation systems. *International journal of scientific and research publications*, 5(1):1–5.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, Online. Association for Computational Linguistics.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2687–2700, Seattle, United States.

Revanth Gangi Reddy, Vikas Yadav, Md Arafat Sultan, Martin Franz, Vittorio Castelli, Heng Ji, and Avirup Sil. 2021. Towards robust neural retrieval models with synthetic pre-training. *arXiv preprint arXiv:2104.07800*.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pages 4334–4343. PMLR.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2022. A thorough examination on zero-shot dense retrieval. *arXiv preprint arXiv:2204.12755*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. No parameter left behind: How distillation and model size affect zero-shot retrieval. *arXiv preprint*.

Rishiraj Saha Roy and Avishek Anand. 2021. Question answering for the curated web: Tasks and methods in qa over knowledge bases and text collections. *Synthesis Lectures onSynthesis Lectures on Information Concepts, Retrieval, and Services*, 13(4):1–194.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. *arXiv preprint arXiv:2204.07496*.

Siamak Shakeri, Cicero dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460.

Xiaoyu Shen. 2022. Deep latent-variable models for text generation. *arXiv preprint arXiv:2203.02055*.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, Bill Byrne, and Adrià de Gispert. 2022a. Product answer generation from heterogeneous sources: A new benchmark and best practices. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 99–110.

Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià de Gispert. 2022b. semipqa: A study on product question answering over semi-structured data. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120.

Xiaoyu Shen and Hui Su. 2018. Towards better variational encoder-decoders in seq2seq tasks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 8155–8156.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019a. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590.

Xiaoyu Shen, Svitlana Vakulenko, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2022c. Low-resource dense retrieval for open-domain question answering: A comprehensive survey. *arXiv preprint arXiv:2208.03197*.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019b. Improving latent alignment in text summarization by generalizing the pointer generator. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3762–3773.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question–context lexical overlap. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 63–72.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2022. Questions are all you need to train a dense passage retriever. *arXiv e-prints*, pages arXiv–2206.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2022. Language models in the loop: Incorporating prompting into weak supervision. *arXiv preprint arXiv:2205.02318*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *ICLR*.

Si Sun, Yingzhuo Qian, Zhenghao Liu, Chenyan Xiong, Kaitao Zhang, Jie Bao, Zhiyuan Liu, and Paul Bennett. 2021. Few-shot text ranking with meta adapted synthetic weak supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5030–5043.

Nandan Thakur, Nils Reimers, and Jimmy Lin. 2022. Domain adaptation for memory-efficient dense retrieval. *arXiv preprint arXiv:2205.11498*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural memoryover symbolic knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3678–3691.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *NAACL*.

Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. 2020. Cross-domain learning for classifying propaganda in online contents. *arXiv preprint arXiv:2011.06844*.

Wenya Wang and Sinno Pan. 2022. Deep inductive logic reasoning for multi-hop reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4999–5009.

Binxia Xu, Siyuan Qiu, Jie Zhang, Yafang Wang, Xiaoyu Shen, and Gerard de Melo. 2020. Data augmentation for multiclass utterance classification–a systematic study. In *Proceedings of the 28th international conference on computational linguistics*, pages 5494–5506.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2022. Laprador: Unsupervised pretrained dense retriever for zero-shot text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3557–3569.

Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Passage ranking with weak supervision. *arXiv preprint arXiv:1905.05910*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021a. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 829–838.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021b. Fine-tuning pretrained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25.

Xiang Yue, Xiaoman Pan, Wenlin Yao, Dian Yu, Dong Yu, and Jianshu Chen. 2022a. C-more: Pretraining to answer open-domain questions by consulting millions of references. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 371–377.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022b. Synthetic question value estimation for domain adaptation of question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1340–1351.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences (2076-3417)*, 10(21).

Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly supervised text classification using supervision signals from a language model. *arXiv preprint arXiv:2205.06604*.

Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Evaluating extrapolation performance of dense retrieval. *arXiv preprint arXiv:2204.11447*.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. Wrench: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective weak supervision for neural information retrieval. In *Proceedings of The Web Conference 2020*, pages 474–485.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé III. 2021. Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9622, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Zhao, Xiaoyu Shen, Wei Bi, and Akiko Aizawa. 2019. Unsupervised rewriter for multi-sentence compression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2240.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, et al. 2022. Hyperlink-induced pretraining for passage retrieval in open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7135–7146.

Dawei Zhu, Xiaoyu Shen, Michael A Hedderich, and Dietrich Klakow. 2022. Meta self-refinement for robust learning with weak supervision. *arXiv preprint arXiv:2205.07290*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021a. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint*.

Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. 2021b. Contrastive learning of user behavior sequence for context-aware document ranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2780–2791.

Shengyao Zhuang and Guido Zuccon. 2022. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. *arXiv preprint arXiv:2204.00716*.

6. A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search

Bibliographic Information

Svitlana Vakulenko, Evangelos Kanoulas, Maarten de Rijke: A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. ACM Trans. Inf. Syst. 39(4): 49:1-49:32 (2021).

Habilitation Candidate CRediT roles

Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, and Project administration.

Copyright Notice

A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search

SVITLANA VAKULENKO, University of Amsterdam, The Netherlands
EVANGELOS KANOULAS, University of Amsterdam, The Netherlands
MAARTEN DE RIJKE, University of Amsterdam & Ahold Delhaize, The Netherlands

Conversational search is a relatively young area of research that aims at automating an information-seeking dialogue. In this paper we help to position it with respect to other research areas within conversational Artificial Intelligence (AI) by analysing the structural properties of an information-seeking dialogue. To this end, we perform a large-scale dialogue analysis of more than 150K transcripts from 16 publicly available dialogue datasets. These datasets were collected to inform different dialogue-based tasks including conversational search. We extract different patterns of mixed initiative from these dialogue transcripts and use them to compare dialogues of different types. Moreover, we contrast the patterns found in information-seeking dialogues that are being used for research purposes with the patterns found in virtual reference interviews that were conducted by professional librarians. The insights we provide (1) establish close relations between conversational search and other conversational AI tasks; and (2) uncover limitations of existing conversational datasets to inform future data collection tasks.

CCS Concepts: • **Information systems → Information retrieval**.

Additional Key Words and Phrases: information-seeking dialogue, mixed initiative, conversational search.

ACM Reference Format:
Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2021. A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. *ACM Transactions on Information Systems* 1, 1, Article 1 (January 2021), 32 pages. https://doi.org/10.1145/3466796

1 INTRODUCTION

Research in conversational AI spans across multiple disciplines, including natural language processing, information retrieval, machine learning and dialogue systems. Its main goal is the development of intelligent conversational systems, which find application across a wide range of domains, such as customer support, education, e-commerce, health, entertainment etc. Several subtasks have been proposed in the context of conversational AI. A recent survey of state-of-art systems in conversational AI groups them according to three main tasks: question answering, task-oriented dialogues, and social chatbots [18]; see Figure 1.

Conversational search has been recognised as an important new frontier within information retrieval [3]. But how does conversational search relate to the tasks previously proposed in the context of conversational AI? Several definitions of conversational search have been proposed to date [3]. All of them agree that the task of conversational search is the development of systems

Authors' addresses: Svitlana Vakulenko, University of Amsterdam, Amsterdam, The Netherlands, s.vakulenko@uva.nl; Evangelos Kanoulas, University of Amsterdam, Amsterdam, The Netherlands, e.kanoulas@uva.nl; Maarten de Rijke, University of Amsterdam & Ahold Delhaize, Amsterdam, The Netherlands, m.derijke@uva.nl.

1046-8188/2021/1-ART1 $15.00
https://doi.org/10.1145/3466796

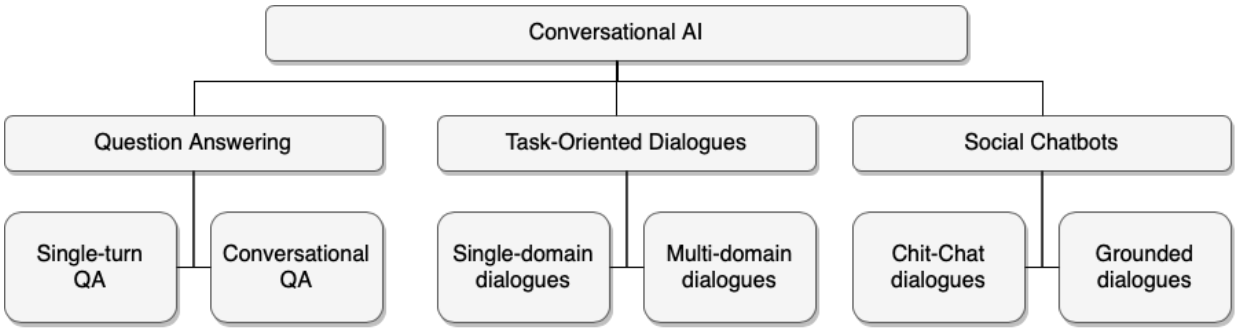


Fig. 1. Tasks within conversational AI as identified by Gao et al. [18]. Where do conversational search and information-seeking dialogues fit in this picture?

that enable information retrieval by means of a natural-language interaction, i.e., dialogue interface [30, 43, 44]. This type of dialogue is referred to as an *information-seeking dialogue*. In this way, conversational search is defined in terms of the type of dialogue it is expected to produce.

Since conversational search is aimed at developing systems that are able to support an information-seeking dialogue, we need to understand the nature and properties of this type of dialogue so as to be able to determine its success. What is an information-seeking dialogue? What are the main characteristics that differentiate it from other dialogue types? Our main goal in this paper is to leverage conversational data for conducting an empirical analysis of information-seeking dialogues collected to date. We aim to gain a better understanding of how information-seeking dialogues differ from dialogues that were collected in the context of other conversational AI tasks, such as chit-chat, conversational QA, and task-oriented dialogue.

In the early days of information retrieval research, reference interviews with a librarian formed an important source of inspiration and information [6, 13, 39]. For the purpose of our study, we collaborated with a global library cooperative (OCLC) to obtain a sample of 560 anonymized virtual reference interviews conducted on-line by professional librarians in 2010. To the best of our knowledge, this is the first systematic study of this scale designed to compare the patterns of interactions observed in virtual reference interviews with the patterns extracted from publicly available dialogue datasets. Our goal was to identify datasets that can best represent the type of interaction that is characteristic of a professional reference interview. The results of such an analysis lead to a recommendation for the datasets that can be used for training and evaluation of conversational search systems designed to mimic this dialogue type.

Our main research questions can be summarised as follows:

(RQ1) What are the structural properties of information-seeking dialogues that differentiate them from other dialogue types?
(RQ2) Which datasets contain dialogues similar to virtual reference interviews conducted by professional librarians?

To answer our research questions we introduce ConversationShape, a framework for dialogue analysis. We focus on the patterns of mixed initiative since the asymmetry of roles was previously identified as the innate property of an information-seeking dialogue, which is due to the knowledge distribution between the conversation participants [8]. Mixed initiative was also identified as one of the core requirements for a conversational search system [30]. Therefore, identifying mixed initiative and describing its use in a successful information-seeking conversation is important for the design and evaluation of conversational search systems.

We demonstrate how our dialogue analysis framework can be applied in practice to conduct a large-scale analysis covering all available conversational datasets. The results of our analysis help

us to define an information-seeking dialogue in terms of a set of measurable properties, and to demonstrate their similarities with, and differences to, other dialogue types.

Our contributions include (1) a large-scale analysis characterising the mix of initiative across different dialogue tasks; and (2) ConversationShape, a methodological framework that has been used to conduct this dialogue analysis and can be re-used to position new dialogue datasets with respect to existing ones.

The remainder of the paper is organised as follows. Section 2 sets the background by providing an overview of existing theoretical foundations for conversational search, previous research focused on dialogue analysis, and mixed-initiative systems. We also review previous studies that manually analyse dialogue transcripts to formulate grounded theories of information-seeking dialogues and mixed initiative in dialogue. We then proceed to describe the ConversationShape analysis framework in Section 3, consisting of fingerprinting, dialogue flow, and asymmetry metrics. Section 4 details our experimental setup and introduces the dialogue datasets that we consider in our analysis. We report the results of applying the ConversationShape framework to these datasets in Section 5, reflect on the outcomes in Section 6, and share our conclusions in Section 7.

2 RELATED WORK

Our work contributes to the body of research devoted to analysing information-seeking dialogues. This type of analysis is useful for informing the theoretical foundations of conversational search by grounding it in empirical observations. Such observations can then be used directly to propose new dialogue models [42, 46].

We start by briefly summarising existing theories of conversational search and information-seeking dialogues. The second subsection provides an overview of the previous research that studies transcripts of information-seeking dialogues and the approaches used to analyse their discourse structure. In the last subsection we review the work focusing specifically on the mix of initiative in dialogues of different types.

2.1 Theories of Conversational Search

Several theories of conversational search have been proposed to date [4, 30, 43]. They are mainly concerned with modeling the set of interactions that occur in the context of an information-seeking dialogue. These ideas can be traced back to the Conversational Roles (COR) Model for generating information-seeking dialogues [37].

Radlinski and Craswell [30] proposed a conversational search model, in which the system interactively provides information to the user and incorporates feedback to elicit user preferences. The goal of the system is to maximise user satisfaction by finding the items with maximal utility. Azzopardi et al. [4] proposed an extended set of user-system actions and subtasks that include suggestion, explanation, navigation, interruption and interrogation.

An important limitation of these theoretical models describing an information-seeking behaviour, in general, is that they are often based on anecdotal evidence drawn from a handful of dialogue transcripts rather than on systematic empirical observations. In contrast to prior work, Trippas et al. [43] performed thematic analysis of the information-seeking dialogue datasets (SCSdata and MISC) by manually labeling 1,710 utterances from 53 dialogue transcripts. Such analysis allowed them to develop a fine-grained labeling schema, SCoSAS, describing the interactions in conversational search. SCoSAS includes 84 labels grouped into three main themes: Task Level, Discourse Lever and Other. Unfortunately, this approach for grounded theory building through dialogue analysis does not scale since it relies on manual annotations of dialogue transcripts. Training a supervised

classifier requires a large number of annotated samples, which are not available on such a fine-grained level. Our goal in this paper is to establish a mechanism that can enable us to continuously refine the theories of conversational search based on the growing volume of empirical data.

There are major gaps in our understanding of the specifics of dialogue interactions that occur in the context of conversational search [40]. In particular, Radlinski and Craswell [30] contribute a fundamental set of requirements for a conversational search system, which includes *mixed initiative*, *user revealment* and *system revealment*. The mechanisms underlying these concepts are yet poorly understood. To fill this gap, we conduct a large-scale dialogue analysis that focuses on several dimensions of mixed initiative in dialogues.

It is important to position conversational search with respect to other research tasks and disciplines. The first Dagstuhl seminar on Conversational Search took an important step in this direction by introducing a typology of conversational search systems [3]; see Figure 2. This typology relates conversational search systems to other types of interactive system, such as chatbots, information retrieval and dialogue systems. In essence, it describes how existing systems can evolve into a conversational search system by gradually extending their capabilities.

The Dagstuhl typology provides an alignment based on the design of existing systems. This perspective has its limitations in biasing the design of a new system towards previously proposed approaches. We take a fundamentally different approach by looking at the expected output of the systems irrespective of their architectural design. To understand the difference between systems we analyse the differences between the dialogues that these systems are designed to produce.

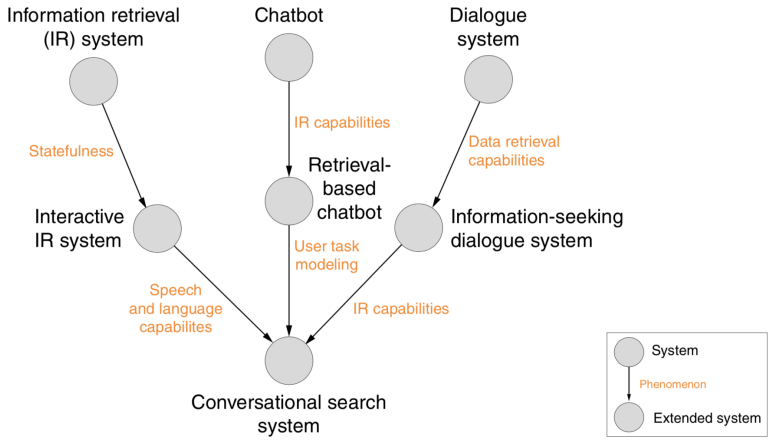


Fig. 2. The Dagstuhl Typology of Conversational Search [3] describes how existing systems can evolve into a conversational search system by dropping different assumptions or limitations.

2.2 Dialogue Analysis

Since the 1980s discourse analysis has been applied to characterise the content of information-seeking dialogues. Belkin et al. [7] suggest that dialogue analysis should be applied to collect patterns from human-human dialogues and identify different information-seeking strategies. These patterns can then serve as reusable scripts within an interactive information retrieval system.

Daniels et al. [13] collect six transcripts of online reference interviews. They use one of the dialogues to derive a set of goals that both intermediary and seeker pursue in a conversation, and use them to annotate utterances in the remaining five dialogue transcripts. Their set of goals for document retrieval interactions include 23 labels grouped into 8 foci, such as Problem Description, Retrieval Strategies, Response Generator, Explanation, etc. In later work, Saracevic et al. [33] perform a similar type of analysis on 40 transcripts of video recordings of reference interviews. They derive

their own classification schema that distinguishes utterances across 8 different categories, such as System explanation, Back channeling, Search tactic, etc.

In more recent work, Trippas et al. [42] reproduce this approach when collecting the Spoken Conversational Search (SCSData) dataset. However, instead of recording interactions with trained intermediaries, such as professional librarians, they recruit volunteers (mostly students) to play the roles of seekers and intermediaries. A similar approach has been employed to collect the MISC corpus [41]. The authors use thematic analysis to derive a new annotation schema (SCoSAS) and annotate both datasets (SCSData and MISC) using this schema [43]. They note important differences in the interaction patterns discovered between the two datasets. These differences are attributed to the difference in task formulation and errors from automated speech recognition.

Next, motivated by the lack of data for training machine learning models, Qu et al. [27] collect forum threads of on-line technical support platforms as samples of information-seeking dialogues. This approach allows the authors to create the large-scale MSDialog corpus with 35.5K dialogues. The authors reuse an existing taxonomy for classifying forum messages with user intents and extend it with four new labels. This taxonomy has been used to annotate a subset of the MSDialog dataset (2K dialogues) using crowdsourcing. Then, the authors extract sequences of utterance labels as dialogue flow patterns and compare them with patterns extracted from the Ubuntu Dialog Corpus [24]. The authors report that while the extracted patterns are the same for both dialogue datasets, their relative frequencies are different.

All of the dialogue analysis studies mentioned above follow the same procedure: (1) derive a set of labels from a sample of the conversational data; (2) annotate utterances with these labels; and (3) extract patterns using these labels, such as label co-occurrences (n-grams). The benefit of this approach is that it allows us to perform different levels of analysis using different annotation schemas. The obvious drawback, however, is the need to manually annotate conversational data; also, the results of the studies are not comparable since they all use different annotation schemas. We follow up on this line of research and show that it is possible to automatically extract and compare patterns of mixed initiative across different dialogue datasets, including but not limited to the SCSData, MISC, MSDialog and Ubuntu Dialog Corpus.

2.3 Mixed-Initiative Design

Mixed-initiative AI systems are specifically designed to enhance human-machine collaboration [12]. In particular, a mixed-initiative dialogue system should be able to recognise the user's cues for initiative switch to provide a response or initiate a discussion when appropriate [11]. Therefore, allocation and transfer of initiative, where initiative is defined as control over the direction of the dialogue flow, is at the core of such systems. Our understanding of mixed initiative in dialogues is important for the design of a conversational search system as a mixed-initiative AI system [30, 40].

State-of-the-art research in the context of mixed-initiative design focuses primarily on the selection/generation of clarifying questions [2, 49]. The need for a clarification typically arises in a situation when the user request is ambiguous, and the system should take an initiative to resolve this ambiguity [38]. Alternatively, the system may also choose to attempt answering the question, even when the user request is ambiguous, and hope to solicit user feedback instead to adjust the answer accordingly. Asking too many or too few questions is likely to result in ineffective and unnatural dialogues. Therefore, we resort to a large-scale dialogue analysis to better understand the strategies chosen by human intermediaries in different situations.

Measuring mixed initiative is essential for assessing dialogue quality [12]. However, the standard dialogue evaluation metrics are scoped to accuracy, relevance and grammaticality of a system response [14, 31]. Our study is designed to fill this gap by introducing a framework for measuring

mixed initiative in dialogues. We demonstrate the utility of this framework by uncovering patterns of initiative across different dialogue types.

The first systematic study of mixed initiative in dialogues is by Walker and Whittaker [47]. They perform a manual analysis of dialogue transcripts and study lexical cues, such as the use of anaphora and different utterance types, as a mechanism for switching control in a dialogue. Their analysis is based on a small sample of dialogues: 24 transcripts from four dialogue datasets across two dialogue types. They use the approach to utterance type classification and the rules for transfer of control between participants has been proposed by Whittaker and Stenton [48], which, in turn, is the earliest investigation into initiative and design of mixed-initiative dialogue systems.

Walker and Whittaker [47] discover different patterns of mixed initiative to be characteristic of different dialogue types. They observe that dialogue participants interact differently depending on the distribution of knowledge between them. In the case of advisory dialogues, such as financial consultation and technical support calls, control over the conversation is shared almost equally between an expert-assistant and an information seeker. In other dialogues, however, an expert-assistant tends to control 90% of the interactions. In these dialogues, experts give out instructions unless interrupted by an information seeker requesting additional clarifications. In this paper, we show how to scale this type of analysis, which has only been performed on a handful of dialogues so far, to thousands of publicly available dialogue transcripts using automated techniques. Our results cast light on the asymmetry of roles that occurs across different dialogue types, including information-seeking dialogues.

In this work, we propose to identify and characterize the general interaction patterns in information-seeking dialogues using the QRFA model [46]. QRFA provides a set of coarse utterance labels that allow for a dialogue analysis on a much higher level of abstraction in comparison with the labels used in other frameworks proposed for modeling the structure of information-seeking dialogues, such as COR [37] and SCS [42].

Our study builds upon the dialogue analysis experiments reported in our previous work [45, 46]. We extend those experiments in several important directions: (1) with an evaluation of the automated utterance classification approach; (2) with a comparison of public dialogue datasets to a sample of virtual reference interviews; and (3) with a comparison of two dialogue analysis approaches introduced in [45, 46] on a large set of dialogue datasets.

3 THE CONVERSATIONSHAPE FRAMEWORK

We propose a dialogue analysis framework, ConversationShape, that helps to detect patterns of mixed initiative in dialogue transcripts. As we will show in our dialogue analysis, these patterns are instrumental in identifying and characterising different dialogue types. ConversationShape includes:

(1) *fingerprinting*, a dialogue representation approach;
(2) a *dialogue flow* model that highlights regular patterns of initiative dynamics; and
(3) a set of *asymmetry metrics* that reflect the distribution of initiative between dialogue participants.

Fingerprinting helps to encode basic structural features of a dialogue and to provide an abstraction suitable for a domain-independent dialogue analysis. We use fingerprints to model dialogue flow and compare dialogue asymmetry across all the datasets listed in Table 3. Both dialogue flow diagrams and asymmetry metrics are methods focused on extracting patterns of mixed initiative in dialogue. Dialogue flow reflects the dynamics of mixed initiative using sequence mining of regular turn switches between the speakers. Asymmetry metrics reflect the balance of initiative distribution along several dimensions of initiative simultaneously. We apply both approaches to analyse the

content of 16 dialogue datasets, thereby revealing similarities and differences in their dialogue structure (see Section 5).

3.1 Fingerprinting Dialogues

To produce a dialogue representation that can be used for both types of dialogue analysis presented in this paper (dialogue flows and asymmetry measurements), we apply fingerprinting to dialogue transcripts. We consider a dialogue transcript to consist of a sequence of utterances $D_i = [U_{i1}, U_{i2}, \ldots, U_{ij}, \ldots, U_{in}]$, where each utterance U_{ij} is a text, i.e., a sequence of characters. Since the original utterances may be long, especially in forum threads (see Table 3), we split them into sentences to simplify utterance classification.

To produce a dialogue fingerprint, we annotate every utterance U_{ij} with the following features:

(1) speaker **role**: Seeker or Assistant;
(2) utterance **length**: an integer for the number of characters in U_{ij};
(3) utterance **type**: Hi, Initiative, NonInitiative, Bye; and
(4) term **repetitions**: a binary vector that indicates the terms in U_{ij} that are repeated, i.e., encountered more than once, within the same dialogue D_i.

Next, we describe these features and the steps that are necessary to produce a dialogue fingerprint. At the end of this section, our fingerprinting approach is illustrated using a sample dialogue from the ReDial dataset [21].

Speaker role. Information-seeking dialogues often have a clearly defined role for each of the dialogue participants. One of the participants has an information need (Seeker) and the other one aims to provide information that can satisfy this need (Assistant).

Utterance length. An important feature that we want to be reflected in the dialogue fingerprint is the share that each of the participants contributes to the dialogue content. We calculate the number of characters in each of the utterances. Below, this will help us to estimate the balance between the dialogue participants in terms of their contributions to the dialogue content. Other metrics can be used for this purpose as well, such as the number of words, subword tokens, phonemes, or the time taken by each of the dialogue participants in case of spoken dialogue.

Utterance type. To recognise utterances that carry initiative in a dialogue, we assign to every utterance U_{ij} one of the labels t from a closed set $T = \{H, I, N, B\}$. Utterance types are assigned independently from the speaker roles. The labels are assigned as follows. Types H for '*Hi*' and B for '*Bye*' are used to filter out utterances that express greetings and farewells. We use type I for *Initiative* to distinguish utterances that include: (1) questions by either of the speaker roles (*Can you help me find ... ? Are you looking for ... ?*); (2) statements containing a request (*Please, help me find ... Please, provide the following information ...*); and (3) statements describing an information need (*I'm looking for ... I need ...*). All other utterances are considered to be of type N.

To assign a label t_{ij} to every utterance U_{ij}, we use a function $F : U_{ij} \mapsto t_{ij}$. This function F is learned by training a supervised classification model. For more information on our approach to training and evaluation of the utterance classifier, see Section 4.2.

Term repetitions. To keep track of repetition patterns, we construct a binary matrix that indicates which frequent term appears in which utterance. A term is considered frequent if it appears in the same dialogue more than once. We start by converting every utterance into a bag-of-words representation and apply standard pre-processing techniques: remove punctuation, split utterances

Table 1. Illustration of the fingerprinting approach for dialogue representation. In this example the dialogue vocabulary consists of two terms that appear more than once in this dialogue: $V_i = \{\textbf{movi}, \textbf{horror}\}$.

Fingerprint					Utterance	Terms
Role	Type	Length	Repetitions			
A	H	4	0	0	Hey!	{hey}
A	I	41	1	0	What kind of movies do you like to watch?	{watch, **movi**}
S	N	42	0	0	I'm really big on indie romance and dramas	{romanc, drama, indi}
A	I	30	1	0	Ok what's your favorite movie?	{favorit, **movi**}
A	I	56	0	0	Staying with that genre, have you seen @88487 or @104253	{genr, stay, 88487, 104253}
A	N	30	0	0	Those are two really good ones	{}
S	N	44	0	1	When I was a kid I liked horror like @181097	{181097, kid, **horror**}
A	N	41	0	0	@Misery is really creepy but really good.	{miseri, creepi}
A	N	32	0	1	I only recently got into horror.	{**horror**}

into words, lowercase, remove stopwords,[1] and apply the English Snowball stemmer. In this manner, every utterance is represented as a set of terms (see Table 1 for an example).

Let us call the set of all frequent terms in dialogue i, the *dialogue vocabulary* V_i. Term repetitions are stored as a binary matrix, where every row j corresponds to the utterance in the dialogue $U_{ij} \in D_i$ and every column k corresponds to the term in the dialogue vocabulary $w_k \in V_i$. Every utterance is encoded into a binary vector using the dialogue vocabulary: $v_{ijk} = 1$ if $w_k \in U_{ij}$ else $v_{ijk} = 0$, where v_{ijk} is the element of the dialogue vocabulary matrix for utterance j and term k of V_i. The dialogue vocabulary can be discarded after all term repetitions are stored in the binary matrix. This approach allows us to easily track repetition patterns in a dialogue: for each term we can identify the speaker who first introduced it and whether it was subsequently reused by another speaker. Term repetitions are part of the dialogue fingerprint.

Definition 3.1 (Dialogue fingerprint)**.** A dialogue fingerprint F_i is a matrix of size $n_i \times m_i$, where n_i is the number of utterances in dialogue i and m_i is the number of features that represent dialogue i. We use the following set of features in our dialogue analysis: $r_{ij} \in R$ is the role a participant plays in a dialogue $R = \{S, A\}$, $t_{ij} \in T$ is one of the utterance types $T = \{H, I, N, B\}$, l_{ij} is the utterance length $l_{ij} = |U_{ij}|$, and $\mathbf{v}_{ij}$ is a vector indicating appearance of the frequent terms in U_{ij}. Thus, each row j of the matrix F_i corresponds to a tuple $\langle r_{ij}, t_{ij}, l_{ij}, \mathbf{v}_{ij} \rangle$ that represents an utterance $U_{ij} \in D_i$.

Illustrative example. Let us consider a sample dialogue to illustrate how fingerprinting works in practice. We will use a snippet of a dialogue transcript from the ReDial dataset [21] (S stands for Seeker, A for Assistant):

(A) Hey! What kind of movies do you like to watch?
(S) I'm really big on indie romance and dramas
(A) Ok what's your favorite movie?
(A) Staying with that genre, have you seen @88487 or @104253
(A) Those are two really good ones
(S) When I was a kid I liked horror like @181097
(A) @Misery is really creepy but really good. I only recently got into horror.

The fingerprint of this dialogue is given in Table 1. Each row corresponds to an utterance annotated with the features described above. Note that we further segment the original utterances provided in the dataset using the sentence boundaries to reduce the utterance lengths and make the classification task, which is required for annotating utterance types, easier. For more details on utterance classification see Section 4.2.

[1] https://raw.githubusercontent.com/stopwords-iso/stopwords-en/master/stopwords-en.txt

Table 2. Mapping from a dialogue fingerprint to QRFA schema reflecting patterns of initiative switch between dialogue participants. Sample utterances were selected from different dialogue datasets using automatic annotations produced by our utterance type classifier.

Role	Type	QRFA	Sample Utterances
Seeker	Initiative	Question	What was the relationship between Coolio and Gangsta's paradise? I am looking for information about the City Centre North B and B hotel. Trying to find a pdf of: Adv Exp Med Biol.
Assistant	Initiative	Request	Which Graphic card is installed on your machine? Would you like to try a different departure station? Is that what you are looking for?
Seeker	NonInitiatiave	Feedback	Yes I like the acting the actors that were in it. Thanks a lot for your recommendations!
Assistant	NonInitiatiave	Answer	I will try to look it up It is a language program that teaches speaking and understanding the language. There are 5 guesthouses that have free parking.

The first column contains the speaker role (A – Assistant, S – Seeker). The second column contains utterance types (H – Hi; I – Initiative; N – NonInitiatiave). The third column contains utterance lengths (measured as the number of characters). The last two columns indicate which frequent terms (dialogue vocabulary $V_i = movi, horror$) appear in which utterance. In this way, the dialogue is encoded into a sequence, where every utterance is represented with exactly 5 features: [A-H-4-0-0, A-I-41-1-0, S-N-42-0-0, …, A-N-32-0-1].

This representation is very compact and privacy-preserving. Since only the structural features of the dialogue are preserved and the dialogue vocabulary is concealed, there is no way to recover the dialogue content from its fingerprint. However, this representation is sufficient for the two types of dialogue analysis that we present in the next two sections.

3.2 Dialogue Flow

A dialogue flow diagram allows us to observe how initiative switches between dialogue participants. We produce a separate diagram for each of the datasets and use them to compare patterns of mixed initiative in information-seeking dialogues and other types of dialogue.

To derive a diagram of dialogue flow, we apply sequence mining to dialogue fingerprints. Since a fingerprint is a matrix we convert it into a single sequence first using utterance types. In this way, every dialogue is represented as a sequence of utterance types.

We focus on the utterance types that indicate dynamics of initiative (Initiative and NonInitiatiave), and extend them to indicate the speaker role as well. The utterances that correspond to greetings (Hello) and farewells (Bye) are ignored. The resulting label set corresponds to the QRFA annotation schema previously used by Vakulenko et al. [46]: two labels for the Seeker role (Query and Feedback) and two labels for the Assistant role (Request and Answer). See Table 2 for the correspondence between our utterance types and the QRFA labels accompanied with sample utterances.

A dialogue flow diagram reflects the counts of all unigrams and bigrams of utterance types across all dialogue sequences in the dataset. The diagrams are constructed using the same template with circles that represent the dialogue start and the dialogue end, and rounded boxes for the QRFA utterance labels. Boxes represent unigrams and arrows represent bigrams. The color intensity (opacity) of the arrows and boxes represents proportions of unigrams and bigrams across all QRFA sequences in the dataset.

The opacity of a box in the diagram corresponds to the respective unigram count normalised by the total count of all unigrams, e.g., the opacity of the Q box is $\frac{\#Q}{\#Q+\#A+\#R+\#F} \times 100\%$. To calculate the opacity for arrows that connect boxes to circles, i.e., the beginning and end of the dialogue, we use bigram counts by prepending < and appending > characters to every sequence, and then normalise them by the number of dialogues, e.g., the opacity of the arrow from the start node to Q is $\frac{\#<Q}{d} \times 100\%$, where d is the number of dialogues in the dataset. The opacity of arrows reflects the proportion of times a sequence begins or ends with the corresponding label.

For example, in the QuAC dataset illustrated in Figure 3 (see Section 4.3), half of the utterances are questions (the opacity of the box labeled Q is 50%) and half of the utterances are answers (the opacity of the box labeled A is 50%). This is unsurprising since QuAC is a conversational QA dataset and the data collection task setup has fixed the dialogue structure in advance. The only possible start of the dialogue is a question by the Seeker (the opacity of the <Q arrow is 100%), and the only possible end of the dialogue is an answer by the Assistant (the opacity of the A> arrow is 100%). The opacities for the unigrams always sum to 100% as do the opacities of the starting arrows and the ending arrows. However, this approach does not work for all other bigram counts.

The same bigram may repeat multiple times in the same dialogue. In this case, the total count will be higher than the number of dialogues. For example, if every dialogue contains two question-answer pairs then the opacity of the QA arrow should be 200%. Alternatively, if we calculate the arrow opacity as the relative proportion of all bigrams (same as for unigrams), most of the arrows will be hardly visible and to tell the difference between their relative proportion will be rather difficult. For example, in a balanced dialogue the fraction of occurrences of each of the eight bigrams QA, AQ, FA, AF, QR, RQ, FR, and RF is equal, $1/8 = 12.5\%$.

To normalise the opacity, we divide every bigram count by the maximum bigram count. The maximum bigram count is obtained by counting all bigrams and then taking the maximum number in this set. For example, if the most common bigram was QA with count 230 the opacity of the QA arrow will be 100% while the opacity of the AQ arrow, assuming its count was 23, will be 10%.

To discard the effect of long turns on the bigram counts (this is especially prominent for forum threads and for transcripts of spoken dialogues), we reduce sequences to a single label per turn. This label indicates whether any of the annotated utterances carries initiative (Q and R) or not (F and A). For example, the fingerprint of the sample dialogue illustrated in Table 1 will result in the following sequence of initiative switches: RFRFA instead of RFR**A**FA. In this way we focus on the initiative switch between dialogue participants and ignore the patterns of initiative within the turn of a single participant. Therefore, there are no arrows between the utterance labels of the same role in the dialogue flow diagrams, i.e., between Q and F, or R and A.

3.3 Asymmetry Metrics

As a complementary analysis to dialogue flow analysis, we propose a set of metrics that reflects other dimensions of initiative beyond utterance types. This set of metrics is designed to measure the distribution of initiative between dialogue participants. We use four metrics to measure initiative in dialogue:

(1) Volume – who talks more in a dialogue?
(2) Direction – who requests information in a dialogue?
(3) Information – who contributes to the dialogue topic? and
(4) Repetition – who follows up on the topic introduced by another participant?

These metrics are derived from the information stored in a dialogue fingerprint. Volume is based on the combination of Role and Length features, Direction is based on the combination of Role and Type, both Information and Repetition metrics are based on the combination of Role and

Repetitions. To compare the metrics between dialogues of different lengths we normalise the scores by the number of utterances, i.e., the number of rows in the dialogue fingerprint n_i.

Volume is associated with explicitly seizing control over the dialogue, effectively turning it into a monologue in extreme cases. We estimate Volume by counting the average number of characters for each of the dialogue participants separately. This is achieved by grouping the values of utterance Length from the dialogue fingerprint by Role and summing them up. For every dialogue D_i the Volume for the role $r \in R$ is computed as:

$$Volume_{ir} = \frac{1}{n_i} \sum_{j=1}^{n_i} l_{ij}[r_{ij} = r]. \tag{1}$$

Direction represents an explicit attempt at controlling the direction of the topic of a dialogue. A question, for example, sets an expectation for another participant to produce a relevant answer. This is achieved by counting the number of utterances with Type *Initiative* grouped by Role. This feature set was used previously for generating dialogue flow diagrams as well. For every dialogue D_i the Direction for the role $r \in R$ is computed as:

$$Direction_{ir} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{I}(t_{ij} = I)[r_{ij} = r]. \tag{2}$$

Information reflects the contribution that a participant makes to the dialogue topic. It is derived from the analysis of the repetition patterns (a binary matrix in the dialogue fingerprint). This metric is motivated by the term frequency counts, which are often used to determine the dialogue topic. We estimate Information by counting the number of frequent terms that were coined by each of the dialogue participants. This is achieved by counting the number of columns in the Repetitions matrix with the first non-zero element pointing to the participant who introduced the term in the dialogue first:

$$Information_{ir} = \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{m_i} \mathcal{I}(v_{ijk} = 1)[r_{ij} = r] \times \mathcal{I}\left(\sum_{g=1}^{j-1} v_{igk} = 0\right). \tag{3}$$

Repetition indicates an explicit follow-up on the topic introduced by another participant. It is measured by counting the non-zero elements in the Repetitions matrix for terms that were not introduced by the speaker. We consider repetition as a type of relevance feedback in the dialogue. Term repetition effectively contributes towards the increase of the term frequency counts. Hence, by repeating the term (following up) a speaker implicitly endorses the contribution of the other dialogue participant and increases its importance with respect to the dialogue topic:

$$Repetition_{ir} = \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^{m_i} \mathcal{I}(v_{ijk} = 1)[r_{ij} = r] \times \mathcal{I}\left(\sum_{g=1}^{j-1} v_{igk}[r_{ij} \neq r] > 0\right). \tag{4}$$

Thus, for every dialogue D_i we measure Volume, Direction, Information and Repetition separately for each of the roles: $Metric_{iA}$ and $Metric_{iS}$, where *Metric* denotes one of the metrics that we have just introduced. Next, we use the average and the difference between $Metric_{iA}$ and $Metric_{iS}$ to compute the averages between all dialogues in the dataset and compare these metrics across different datasets.

The difference between $Metric_{iA}$ and $Metric_{iS}$ allows us to compare the distribution (that is, the balance of initiative) between the dialogue participants. To produce a score in the range $[-1; 1]$, with 0 indicating the exact balance of initiative, 1 and -1 indicating dominance of initiative by

either Assistant or Seeker, respectively, we calculate $\Delta Metric$ across all d dialogues in the dataset as follows:

$$\Delta Metric = \frac{1}{d} \sum_{i=1}^{d} \frac{Metric_{iA} - Metric_{iS}}{Metric_{iA} + Metric_{iS}}. \tag{5}$$

Illustrative example. We show how the asymmetry metrics are computed using the sample dialogue fingerprint from Table 1. The Assistant clearly dominates the dialogue by the number of utterances and their relative length. This also leads to a difference between the number of characters ascribed to each of the dialogue participants, which is reflected in the $\Delta Volume$ metric:

$$Volume_{iA} = \frac{4 + 41 + 30 + 56 + 30 + 41 + 32}{9} \approx 42 \tag{6}$$

$$Volume_{iS} = \frac{42 + 44}{9} \approx 10 \tag{7}$$

$$\Delta Volume = \frac{42 - 10}{42 + 10} \approx 0.62. \tag{8}$$

Moreover, all the questions in this dialogue originate from the Assistant. This information is reflected in $\Delta Direction$ metric:

$$Direction_{iA} = \frac{3}{9} \approx 0.33 \tag{9}$$

$$Direction_{iS} = 0 \tag{10}$$

$$\Delta Direction = \frac{0.33 - 0}{0.33 + 0} = 1. \tag{11}$$

Both dialogue participants contribute to the dialogue topic. The Assistant was the first to introduce *movies* as the dialogue topic. Later, the Seeker introduced *horror* as a subtopic:

$$Information_{iA} = Information_{iS} = \frac{1}{9} \approx 0.11 \tag{12}$$

$$\Delta Information = \frac{0.11 - 0.11}{0.11 + 0.11} = 0. \tag{13}$$

The Assistant followed up on the topic introduced by the Seeker, also repeating the word *horror* that was introduced by the Seeker. Thereby, the Assistant demonstrates the ability to lead the dialogue as well as the readiness to follow up on the topic of interest introduced by the Seeker:

$$Repetition_{iA} = \frac{1}{9} \approx 0.11 \tag{14}$$

$$Repetition_{iS} = 0 \tag{15}$$

$$\Delta Repetition = \frac{0.11 - 0}{0.11 + 0} = 1. \tag{16}$$

4 EXPERIMENTAL SETUP

In this section we describe the setup used for our large-scale dialogue analysis. We list all the datasets considered in our analysis and point out the relations between them that are known a priori. In this section we also provide details on training and evaluation of the utterance classification model, which we use to annotate utterance types in all datasets as part of the dialogue fingerprints.

4.1 Dialogue Datasets

Our analysis is focused on the information-seeking dialogue datasets that have previously been introduced or analysed in the context of conversational search. We complement them with datasets for other dialogue types indicated in Figure 1: conversational QA, task-oriented, grounded and chit-chat dialogues.

Table 3 groups the datasets by type and summarises their main characteristics. We differentiate dialogues by domain, modality (text or speech), source (forum, chat, task or book), and setup (natural or simulated information need). Table 3 also provides basic statistics for each of the datasets in terms of the number of dialogues, the dialogue and utterance lengths, and the average number of utterances per turn.

Most of the datasets have been collected in an artificial setting (see the value *simulated* under Setup in Table 3), i.e., the participants were instructed to interact with each other or with a chatbot (Meena/Mitsuku and Meena/Meena). These interactions are usually mediated by a text-based chat interface (see the value *text* for Modality and *chat* for Source in Table 3).

4.1.1 Information-seeking dialogues. There is a subset of real information-seeking dialogues that occurred on-line without intrusion of researchers (see Setup *natural* in Table 3). These datasets have either been extracted from Q&A forums or from on-line chatrooms. We refer to this subset of information-seeking dialogue datasets collected in natural settings as IN, and the rest of the information-seeking dialogue datasets collected in the simulated setting as IS.

Q&A forums (IN). MSDialog [27] has been collected on a technical support forum provided by Microsoft, and MANTIS [26] is from the community question-answering portal Stack Exchange. As is evident from Table 3, both datasets have very different basic statistics from the other dialogue datasets: forum threads are usually much shorter than text- and speech-based dialogues and their utterances are much longer.

On-line chats (IN). Unlike forums, on-line chatrooms provide an opportunity for real-time synchronous communication, which enables a more dynamic interaction that more closely resembles a human dialogue. In particular, a reference interview with a librarian, which is a classic example of an information-seeking dialogue, has been replaced with a virtual reference interview that occurs in a private chatroom [28].

We obtained a sample of dialogue transcripts of virtual reference interviews from OCLC, which is a non-profit global library cooperative. This is a random sample with 560 anonymised transcripts of virtual reference interviews, which took place from June 2010 through December 2010. The OCLC dataset contains real information-seeking dialogues mediated by professional librarians, who were trained to provide reference services.

The most similar dataset of this type that is publicly available, is the Ubuntu Dialogue Corpus [24]. It contains dialogue transcripts extracted from a public chat of the Ubuntu user community. The goal of this chat is to provide technical support and advice on software-related issues. Note that the dialogues in this dataset are more informal than those in the OCLC dataset since they occur between peers rather than between a customer and a professional service provider.

Spoken Conversational Search (IS). The SCSdata [42] and MISC [41] datasets resulted from two separate laboratory data collection tasks, in which pairs of human volunteers interacted in the context of an information-seeking task. One of the volunteers had access to the textual description of the information need, and the other one was using a search engine to find relevant information that can satisfy this need. Their spoken dialogues were recorded and then transcribed, either manually in the case of SCSdata or automatically in the case of MISC.

Conversational recommendation (IS) shares a number of similarities with the conversational search task and the boundaries between the two are not explicitly defined [17, 19]. Thus, we consider

dialogues collected to inform the conversational recommendation task to constitute information-seeking dialogues. Both datasets we consider in the context of conversational recommendation, ReDial [21] and CCPE [29], focus on the movie recommendation scenario. While ReDial contains dialogues in which human participants were instructed to recommend movies to watch, CCPE is focused specifically on the preference elicitation phase, in which the assistant is instructed to learn more about the user tastes rather than make a recommendation.

Conversational QA (IS) datasets result from task designs that fix the structure of an information-seeking conversation in advance, following a pre-defined template. Therefore, such datasets are not suitable for discovering and analysing the structure of naturally occurring dialogues. For example, the QuAC dataset [10] consists of dialogues that contain sequences of question-answer pairs. The Qulac dataset [2] contains synthetic dialogues of length 3, in which every user question is followed by a clarifying question and a user response. Since the utterance types are known a priori, we use this dataset to train our utterance classification model.

4.1.2 Other dialogue types. The set of information-seeking dialogues is complemented with other dialogue datasets collected to train and evaluate models for conversational QA, task-oriented, grounded and chit-chat dialogues.

Task-oriented (TO) dialogues aim to accomplish a certain task or a sequence of related subtasks, such as booking a restaurant and calling a cab. We include the MultiWOZ dataset [9] in our analysis to discover the typical structure of a task-oriented dialogue and compare it with the structure of an information-seeking dialogue. MultiWOZ is a large-scale dataset that spans across multiple domains and topics. While we include a single task-oriented dataset in our analysis, we consider it representative of this dialogue type due to its size (10K dialogues is an order of magnitude larger than all previous annotated task-oriented corpora [9]) and diversity (single-domain and multi-domain dialogues about restaurants, hotels, attractions, taxi, trains, hospitals and police).

Chit-chat (CC) dialogues are used to design and evaluate social chatbots [18]. The primary goal of a social chatbot is to entertain the user by holding a human-like conversation. The Meena dataset consists of three subsets (Meena/Mitsuku, Meena/Meena and Meena/Human) used for chatbot evaluation [1]. Two of them (Meena/Mitsuku and Meena/Meena) contain transcripts of human-machine dialogues produced by volunteers interacting with two social chatbots. The third subset (Meena/Human) contains transcripts of human-human dialogues that follow the same setup and are used as a reference point for chatbot evaluation. The DailyDialog dataset [22] contains samples of dialogues that occur in common daily situations, such as shopping, doctor appointment etc. It is different from all other datasets listed in Table 3 since these dialogues were extracted from textbooks for English learners. This dataset is commonly considered as chit-chat [35].

Knowledge-grounded (KG) dialogues are similar to chit-chat but the additional goal is also to communicate certain information (knowledge). This knowledge is provided as input (grounding) to the dialogue model either as text (PersonaH [34] and WoW [15]) or a knowledge graph (OpenDialKG [25]). For example, the PersonaH dataset contains human-human dialogues grounded in short text snippets. These snippets contain descriptions of personal information, such as hobby and occupation, that the participants are supposed to use in their replies [34]. Every dialogue in the WoW dataset revolves around a topic discussed in one of the Wikipedia articles and dialogues in the OpenDialKG dataset use facts stored in the Freebase knowledge graph.

All information-seeking datasets have turns annotated with the participant roles, except for the Ubuntu dataset. Since the dialogues were automatically extracted from a public chatroom, the turns in Ubuntu are annotated only with the user handles. We use a simple heuristic and assume the first speaker, who initiates the dialogue, to be the Seeker. This decision is motivated by the observation that the Seeker is more likely to initiate the conversation to indicate the information

Table 3. 16 dialogue datasets used in our analysis: 15 public and 1 private (OCLC). The Meena dataset consists of three subsets, which contain human-human and human-machine dialogues (*) that were collected for a chatbot evaluation. All other dialogue datasets contain only human-human dialogues. Dial. len. – average dialogue length calculated as the number of utterances. Utt. len. – average utterance length calculated as the number of words. Utt./turn – average number of utterances per single dialogue turn produced by one of the dialogue participants. The number in brackets is a standard deviation from the mean. Outliers for the different dimensions are highlighted in **bold**.

Dataset	Dialogues	Domain	Type	Subtype	Modality	Source	Setup	Dial. len.	Utt. len.	Utt./turn
MSDialog [27]	35,500	tech	info-seek	IN	text	forum	**natural**	9 (25)	**67 (87)**	1 (3)
MANTIS [26]	1,400	multi	info-seek	IN	text	forum	**natural**	4 (1)	**98 (160)**	1 (0)
OCLC[a]	560	library	info-seek	IN	text	chat	**natural**	25 (20)	12 (15)	1 (1)
Ubuntu [24]	1,200,000	tech	info-seek	IN	text	chat	**natural**	6 (8)	10 (9)	1 (1)
SCSdata [42]	37	web	info-seek	IS	**speech**	chat	simulated	27 (21)	16 (25)	1 (0)
MISC [41]	110	web	info-seek	IS	**speech**	chat	simulated	**120 (47)**	7 (7)	**2 (2)**
ReDial [21]	10,000	movies	info-seek	IS	text	chat	simulated	18 (5)	6 (5)	1 (0)
CCPE [29]	502	movies	info-seek	IS	text	chat	simulated	23 (6)	12 (13)	1 (0)
Qulac [2]	10,277	web	info-seek	IS	text	**task**[b]	simulated	3 (0)	8 (3)	1 (0)
QuAC [10]	11,600	Wikipedia	info-seek	IS	text	chat	simulated	14 (4)	9 (7)	1 (0)
MultiWOZ [9]	10,000	multi	**task-orient**	TO	text	chat	simulated	13 (5)	13 (6)	1 (0)
Meena/Mitsuku* [1]	100	open	social	CC	text	chat	simulated	18 (4)	8 (10)	1 (0)
Meena/Meena* [1]	91	open	social	CC	text	chat	simulated	19 (5)	6 (4)	1 (0)
Meena/Human [1]	95	open	social	CC	text	chat	simulated	15 (2)	13 (10)	1 (0)
DailyDialog [22]	11,000	multi	social	CC	text	**book**[c]	simulated	7 (4)	13 (10)	1 (0)
PersonaH [34]	102	personal	social	KG	text	chat	simulated	12 (0)	9 (4)	1 (0)
WoW [15]	22,000	Wikipedia	social	KG	text	chat	simulated	9 (1)	16 (7)	1 (0)
OpenDialKG [25]	13,800	Freebase	social	KG	text	chat	simulated	6 (2)	12 (6)	1 (0)

[a]The dataset is an intellectual property of OCLC https://www.oclc.org and is subject to a licence agreement.
[b]In Qulac, participants were not paired to participate in a live conversation but added their responses into an on-line form given the previous utterance and the information need description as a prompt.
[c]These dialogues were extracted from a textbook for English learners, i.e., likely created by a single author.

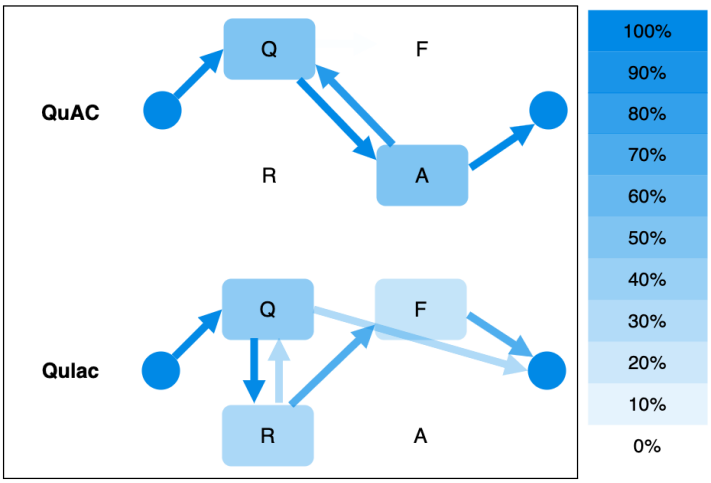


Fig. 3. Models of dialogue flow extracted from QuAC and Qulac datasets using automatically annotated utterance labels. The vertical bar indicates the correspondence between the opacity levels for boxes/arrows and the unigram/bigram frequencies, respectively. This color coding schema is consistent across all the dialogue flow models in the following sections.

need. All dialogues in OCLC and MSDialog are also initiated by the Seeker. We apply the same heuristic to assign the roles in social dialogues to show that it provides for a random assignment if the participant roles are symmetric.

4.2 Utterance Classification Model

We train our utterance classification model by fine-tuning a pre-trained RoBERTa base with 12 hidden layers and 12 attention heads [23]. We use a combination of four datasets to train our model: QuAC, SPAADIA [20], Qulac and NPS chat [16].

QuAC contains question-answer pairs, which we use as samples for Inititative (I) and NonInitiative (N) utterance types, respectively. From the NPS chat corpus, we obtain additional examples of questions (I), greetings (H) and farewells (B).

Qulac and SPAADIA are used to train the classifier to recognise those information requests that are not formulated as questions. Every sample in Qulac contains a topic description (e.g., “Find a dieting advice”), a facet description (e.g., “Find crash diet plans”), a clarifying question (e.g., “do you want to know if dieting is safe”), and an answer (e.g., “no I don’t.”) We use questions, topic and facet descriptions to train I, and answers to train N type.

The SPAADIA dataset provides examples of the four main sentence types: declarative (statements), imperative (commands), exclamatory (exclamations) and interrogative (questions). We use declarative sentences to train the N type; imperative and interrogative sentences for the I type.

The resulting dataset contains 86K utterance-labels pairs (H: 1.4K, I: 42K, N: 42K, B: 195). We randomly split it into two subsets: 77.5K for training and 8.6K for testing our classifier.

4.3 Evaluation of Utterance Classification

Our utterance classification approach achieves a macro-average F1 score of 0.942 on the held-out test set. Then, we use this classification model to annotate utterances in all the datasets listed in Table 3. The Ubuntu dataset is too large even for an automated utterance classification (more than 1M dialogues). To save computational resources, we annotated a subset of the first 50K dialogues and consider it to serve as a representative sample for our dialogue analysis.

We performed an error analysis by manually inspecting the results. Two of the paper authors labeled 1,494 randomly selected utterances independently. We made sure to keep the balance of

classes in the random sample we selected (half was labeled as Initiative by our classifier and half as NonInitiative). There were only 12 cases of disagreement, in total (99% utterances received the same label by both annotators).

The accuracy of our utterance classifier on this random subset is 92% (113 utterances were classified incorrectly). Here are some examples of utterances that were labeled as Initiative by our utterance classifier, which demonstrate that the classification model also learned to recognise different formulations for information requests beyond questions: (1) "*I have found the files, but what am I supposed to use to open them?*" (2) "*Tell me about your favorite movie.*" (3) "*I'm looking for some suggestions for good movies.*"

To verify the impact of the utterance classification on the results of the dialogue flow analysis, we produced dialogue flow diagrams for the QuAC and Qulac datasets using automatic annotations (see Figure 3). While the dialogue flow diagram for the QuAC dataset perfectly fits our assumption about these dialogues as a sequence of question-answer pairs, the triples in the Qulac dataset turned out to be more diverse than we initially thought. In some cases (31% of the dialogues) the Seeker replied to a clarifying question with feedback paired with a follow-up question. Examples: "yes specifically how is it different from hdl and vldl", "yes and also how would i apply", "can you just show me the human society's homepage." By manually examining such cases, we conclude that our utterance classifier successfully learned to distinguish questions erroneously annotated as NonInitiative statements in the training set, i.e., the classifier learned to assign correct labels with overfitting the training data.

To extend the evaluation of our utterance classifier to other dialogue datasets that were not used during training, we utilise the manual labels provided along with the SCSdata and MSDialog-Intent datasets. MSDialog-Intent is a subset of MSDialog with 2,199 dialogues with utterances manually annotated by crowd workers.

SCSdata and MSDialog-Intent were annotated with two disjoint label sets. Utterances in SCSdata were manually annotated with 83 labels, called actions, such as "Info about document", "Query repeat", "Performance feedback", etc. The MSDialog annotators used a set of 12 labels, such as "Original Question", "Clarifying Question", "Potential Answer" and "Positive Feedback". Therefore, we had to manually map both label sets to the QRFA labels.

Table 4 shows the mapping that we established between the manually annotated labels and the QRFA schema. The table also contains examples of utterances extracted from both of the datasets for each of the QRFA labels.

Our utterance labelling approach achieved a micro-average F1 score of 0.8 on the SCSdata dataset and only 0.17 on the MSDialog-Intent dataset. 84% of the errors for the SCSdata dataset are due to the inability of the model to recognise Initiative. In contrast, 70% of the errors on the MSDialog-Intent dataset are due to the model predicting Initiative where it was not annotated by the crowd workers.

By manually examining all errors in the SCSdata dataset and a random sample of errors from the MSDialog-Intent dataset, we identify two main reasons behind these misclassification results: (1) utterances annotated as Initiative do not contain an explicit question (see "Original Question", "Intent clarification" and "Asks to repeat" in Table 4); and (2) utterances that contain explicit questions or requests were not annotated as such (see "Further Details", "Potential Answer", "SERP with modification + Interpretation" in Table 4). In particular, most of the utterances produced by the Assistant in MSDialog-Intent contain a request for feedback, which was not annotated by the crowd workers.

We observed that the annotation schemas could not be directly mapped in some cases (see different examples of "Intent clarification" in Table 4, which may be Initiative as well as NonInitiative). Designing a guideline for the manual annotation task is challenging and error-prone (see questions

and requests highlighted in Table 4 but ignored by the human annotators). Our automated utterance classifier is based on shallow features and designed to discriminate only between sentence types (interrogative and imperative versus declarative sentences). In this evaluation, we showed that the results of our classifier not only correlate with the human judgement of utterance type but can be also used to infer missing labels.

Figure 4 shows the dialogue flow diagrams extracted from the SCSdata and MSDialog-Intent datasets:

(1) manually annotated by the dataset authors (SCSdata*) or crowdsourced (MSDialog-Intent*);
(2) automatically annotated with our utterance classifier (SCSdata and MSDialog-Intent).

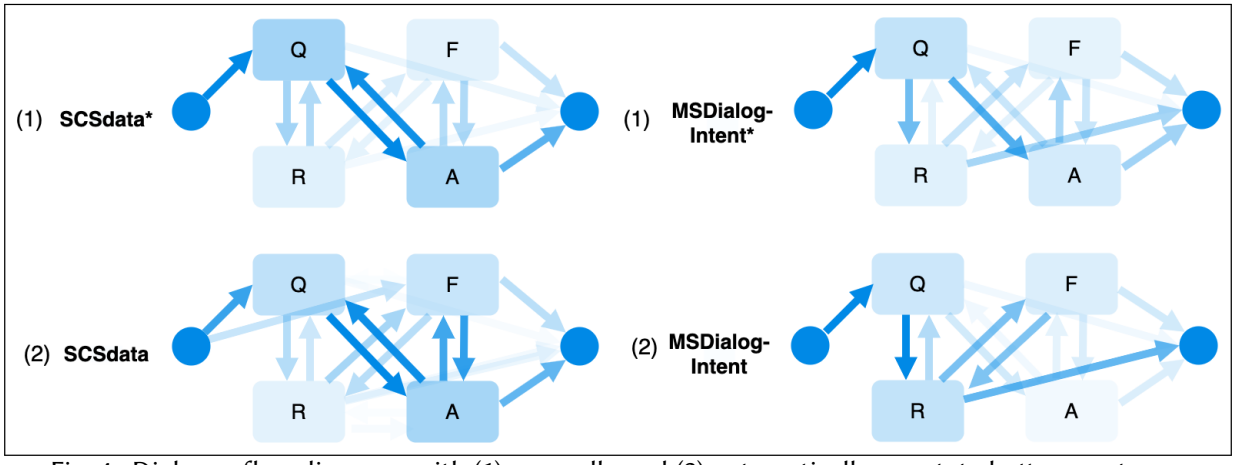


Fig. 4. Dialogue flow diagrams with (1) manually and (2) automatically annotated utterance types.

5 RESULTS

In this section, we compare the dialogue datasets introduced in Section 4.1 using the dialogue flow diagrams and asymmetry metrics. We start with the dialogue flow analysis and use it to compare information-seeking dialogues with other dialogue types.

Dialogue flow diagrams provide a good overview of dialogue datasets but it is not easy to use them to compare, or to quantify similarities between, dialogues. We show how the results obtained from the dialogue flow analysis can be further extended by adding other dimensions of mixed initiative offered by the asymmetry metrics. The asymmetry metrics allow us to represent every dialogue as a vector and embed them into a common vector space. We show that this approach makes it easier to compare dialogues along alternative dimensions and retrieve similar dialogues.

5.1 Dialogue Flow Analysis Results

The dialogue flow diagrams produced for information-seeking dialogues are presented in Figure 5. They are grouped into natural (IN, on the left) and simulated dialogues (IS, on the right).

The first thing to notice is that the diagrams vary a lot. Therefore, we conclude that there are different subtypes of information-seeking dialogues according to the patterns of initiative. This observation motivates us to reconsider the existing dialogue classification schema presented in Table 3 and establish a new classification based on the patterns of mixed initiative.

5.1.1 Search-Support dialogue classification. Most of the dialogues in Figure 5 demonstrate asymmetry of initiative towards one of the roles. We refer to the dialogues in which the Seeker asks most of the questions and the Assistant provides most of the answers ($QA > RF$), as *Search* dialogues. For examples of Search dialogues in Figure 5, see MANTIS, Ubuntu, and SCSdata.

Table 4. Mapping between MSDialog-Intent and SCSdata annotations to the QRFA labels. The sample utterances are shortened due to their length. *Cursive* indicates examples of misclassification, in which information requests were not manually annotated in the original datasets.

Dataset	Original label	QRFA	Sample utterance
MSDialog-Intent	Original Question	Q	Hello, I have done a reset ... I get to the same error. What do I do next?
	Original Question	Q	I called ... Note that I never ... which I cant.
	Information Request	R	What is the model of the computer? Have you tried ...?
	Follow Up + Repeat Question	AR	Hey there, just did a factory reset... Did you find a solution in the end ...?
	Further Details	F	no updates available I know of-*please send me a list of any ...*
	Greetings/Gratitude + Positive Feedback	A	Hi... Thank you for posting back with the result. Glad to know the issue resolved. *Feel free to post us if you need any assistance ...*
	Potential Answer	A	Once that you've restarted your computer, we suggest that you ... *We would like to know if there is an antivirus software installed on your computer. We look forward for your response.*
SCSdata	Initial information request	Q	In which countries... in which European countries do they grow cinnamon?
	Access source	Q	Can you just look at the news dot com
	Intent clarification	Q	Oh I'm I'm looking to find out uhm what what jobs are being outsourced from the US specifically to India
	Intent clarification	Q	Yes that's that's twenty to twenty-four
	Asks to repeat	R	Passenger and
	Asks to repeat	R	Oh per person
	Asks to repeat	R	Sorry
	Feedback on what is happening	F	So I'll ask a second question
	SERP with modification + Interpretation	A	It just says a lot of comparing and uhm like there are some articles that start to talk about like uhm sort of plants and stuff
	SERP with modification + Interpretation	A	The next one is the impact ... *are you wanting anything newer?*

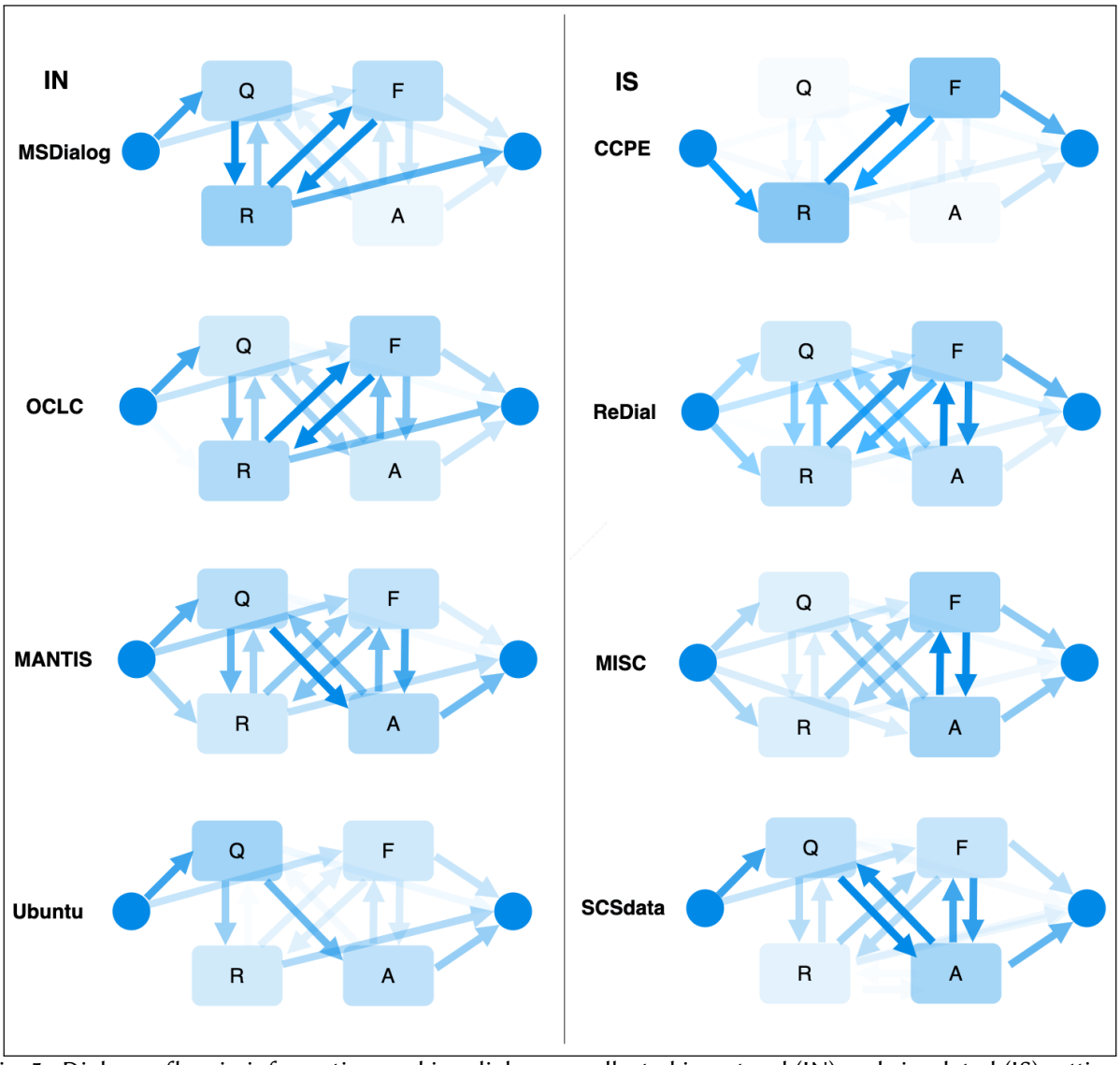


Fig. 5. Dialogue flow in information-seeking dialogues collected in natural (IN) and simulated (IS) settings.

Search dialogues are contrasted with *Support* dialogues, in which the Assistant plays a more active role by asking more questions and requesting information from the Seeker ($RF > QA$). For examples of Support dialogues in Figure 5, see MSDialog, OCLC, CCPE and ReDial. In CCPE almost all questions were asked by the Assistants since it is a preference elicitation dataset. In comparison, ReDial has a smaller difference between the roles (82% RF versus 56% QA bigrams).

5.1.2 Other dialogue types. For comparison we also produced dialogue flow diagrams for other dialogue types than information-seeking dialogues. All dialogues are grouped by following the Search-Support classification criterion introduced in Section 5.1.1:

$$QA > RF \rightarrow Search \tag{17}$$

$$QA < RF \rightarrow Support \tag{18}$$

We found that *Search* dialogues turn up very often among other dialogue types, including knowledge-grounded and chit-chat conversations (see Figure 6). These dialogues exhibit similar interaction patterns as the ones observed in information-seeking dialogues extracted from on-line community discussions (MANTIS and Ubuntu).

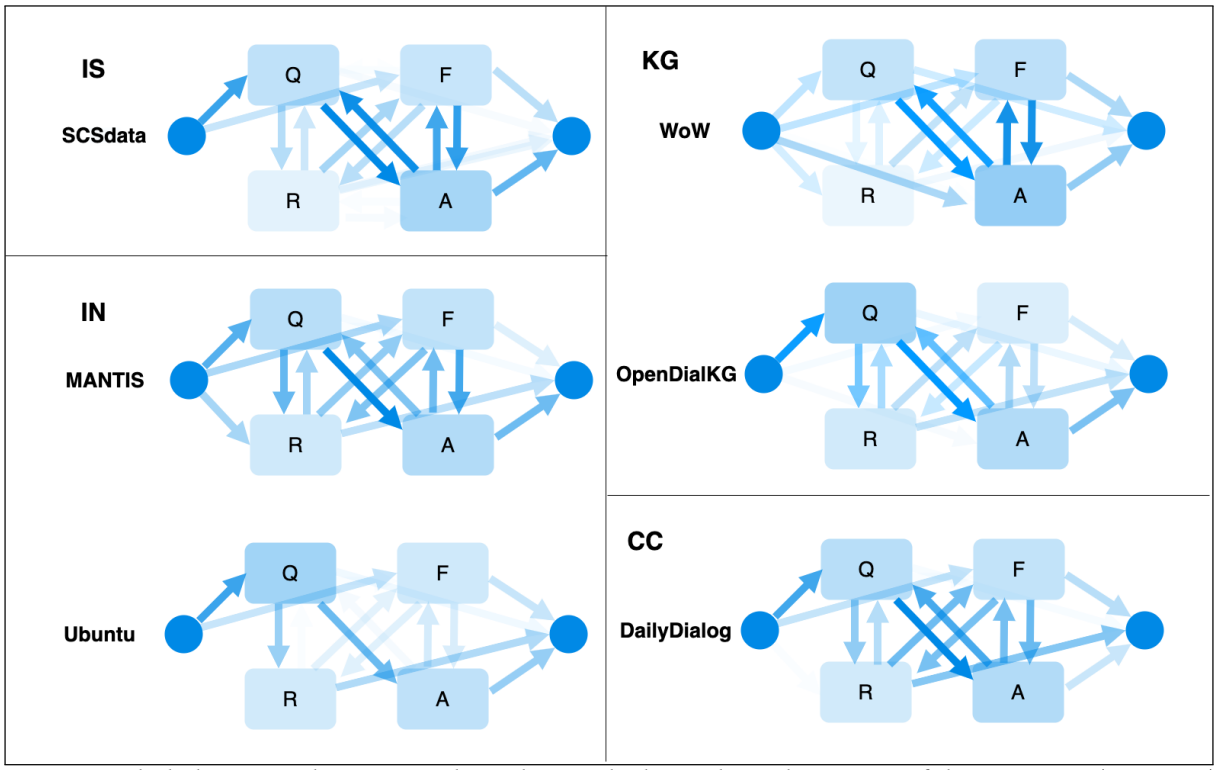


Fig. 6. Search dialogues with initiative skewed towards the Seeker asking most of the questions ($QA > RF$).

Dialogues classified as Support (both information-seeking and non information-seeking) are shown in Figure 7. Dialogues with this interaction pattern are produced either (1) by professional intermediaries in on-line forums and chat-rooms; (2) in a conversational recommendation setting; or (3) in a task-oriented dialogue setting. It is also clear from Figure 7 that none of the simulated datasets mirrors the patterns of initiative discovered in naturally occurring dialogues (MSDialog and OCLC). ReDial has too much chit-chat (FA-AF) and MultiWOZ has too many questions by the Seeker (QR-RQ).

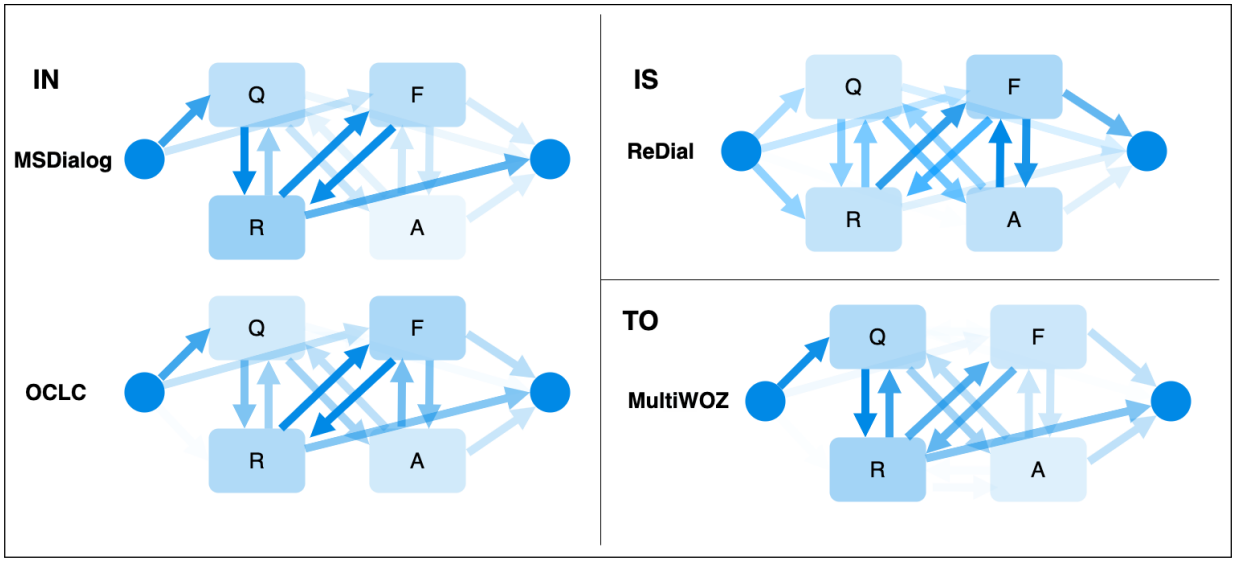


Fig. 7. Support dialogues with initiative skewed towards the Assistant asking most of the questions (QA<RF)

Finally, in Figure 8 we show the dialogue flow diagrams for the datasets that do not fall into the previous two categories, i.e., in which $QA \approx RF$. Notice how symmetry of speaker roles in social dialogues causes the diagrams to be symmetric as well. In Meena and PersonaH datasets the speakers were instructed to have a casual chat without any specific roles assigned. We refer to these dialogues as *information-sharing* dialogues, in contrast with information-seeking dialogues.

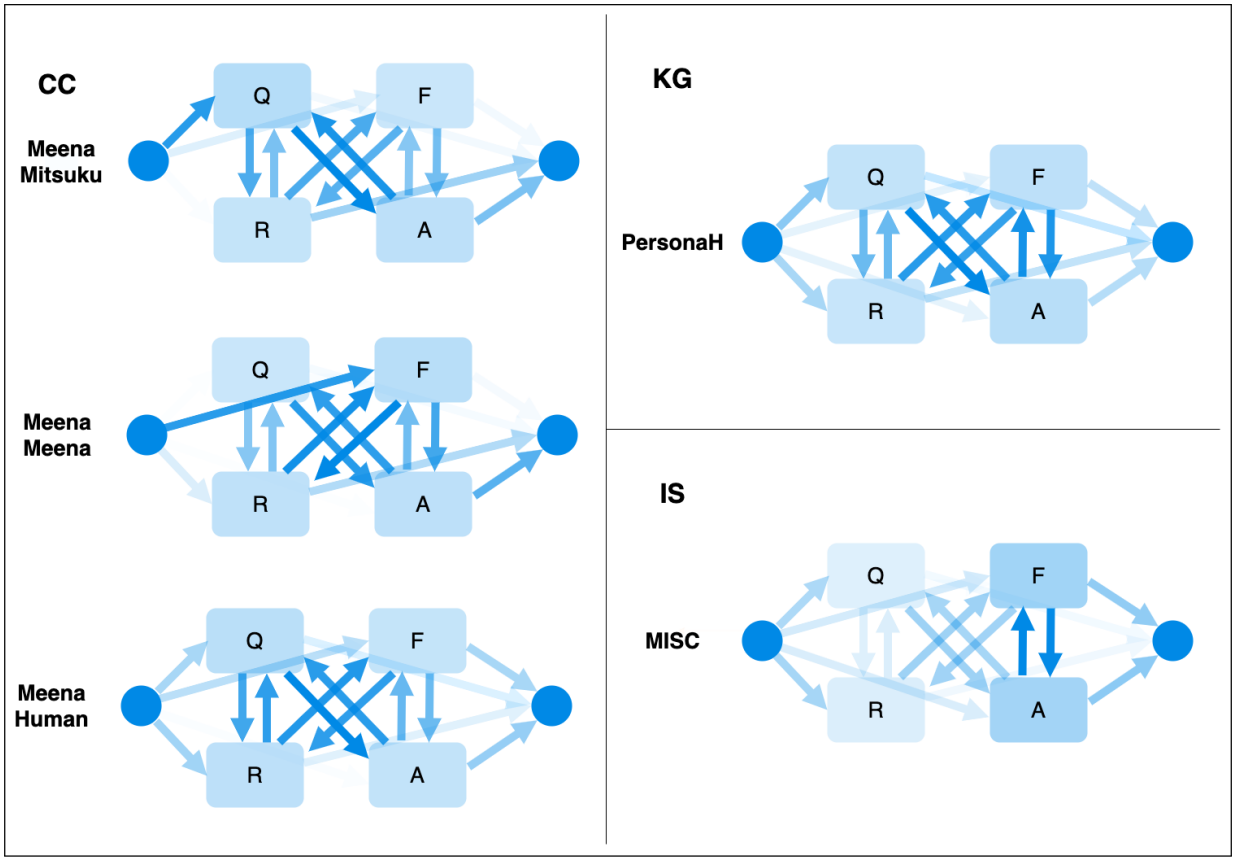


Fig. 8. Information sharing dialogues appear symmetric due to equal distribution of initiative ($QA \approx RF$).

We established the differences between dialogues using the dialogue flow diagrams that cannot be explained by the initial dialogue classification provided in Table 3. Figures 6-8 demonstrate that dialogues of different types often have similar structure and dialogues of the same type may exhibit very different structural patterns. Therefore, we proposed to group similar dialogue datasets based on their structural patterns into three new dialogue types: Search, Support and Sharing.

5.2 Analysis Result of the Asymmetry Metrics

The results of embedding dialogue datasets using asymmetry metrics are displayed in Figures 9–11. In all plots, information sharing dialogues appear close to the origin of the coordinates, which represents a balance of initiative. Information-seeking dialogues can be characterised by different types of asymmetries and are located away from the origin.

Volume – who talks more in a dialogue? In the majority of dialogue datasets that we consider in our analysis, the Assistant talks more than the Seeker. There are only three datasets where the Seeker talks much more than the Assistant: CCPE, Qulac and Ubuntu ($\Delta Volume < -0.25$). The Ubuntu dataset is an outlier among the information-seeking dialogue datasets because it contains many samples where the questions posed by the Seekers were not answered.

Table 5. Refined dataset types based on dimensions of mixed initiative identified via data flow diagrams and asymmetry metrics. The datasets are ordered by $\Delta Direction$. The negative values for all dimensions are highlighted in bold.

Dataset	Type	Our Type	ΔVolume	ΔDirection	ΔInformation	ΔRepetition
CCPE	Inf.seek(simulated)	Support	**-0.35**	0.87	**-0.17**	**-0.59**
MSDialog	Inf.seek(natural)	Support	0.20	0.35	**-0.11**	0.70
OCLC	Inf.seek(natural)	Support	0.37	0.30	0.06	0.58
ReDial	Inf.seek(simulated)	Support	**-0.03**	0.12	0.03	**-0.09**
Meena/Meena	Chit-chat	Sharing	0.09	0.09	**-0.16**	0.32
MultiWoZ	Task-oriented	Support	0.16	0.05	**-0.31**	0.53
MISC	Inf.seek(simulated)	Sharing	0.01	0.00	0.04	**-0.02**
Meena/Human	Chit-chat	Sharing	0.02	**-0.03**	**-0.03**	0.08
Meena/Mitsuku	Chit-chat	Sharing	0.29	**-0.05**	0.02	0.09
PersonaH	Knowledge-grounded	Sharing	**-0.01**	**-0.06**	**-0.02**	0.10
Qulac	Inf.seek(simulated)	Search	**-0.30**	**-0.10**	**-0.83**	0.60
DailyDialog	Chit-chat	Search	**-0.04**	**-0.24**	**-0.16**	0.16
MANTIS	Inf.seek(natural)	Search	0.07	**-0.24**	**-0.30**	0.66
Ubuntu	Inf.seek(natural)	Search	**-0.33**	**-0.39**	**-0.41**	0.31
SCSdata	Inf.seek(simulated)	Search	0.27	**-0.42**	0.10	0.42
OpenDialKG	Knowledge-grounded	Search	0.01	**-0.44**	0.00	0.08
WoW	Knowledge-grounded	Search	0.16	**-0.45**	0.20	**-0.07**
QuAC	Inf.seek(simulated)	Search	0.33	**-1.00**	0.35	0.06

Direction – who requests information in a dialogue? The Seeker tends to ask more questions in the majority of datasets. This does not hold for CCPE, MSDialog, OCLC, ReDial, Meena and MultiWoZ. CCPE is the major outlier since most of the questions are asked by the Assistant to elicit user preferences.

Information – who contributes to the dialogue topic? The scatter plot in Figure 10 shows that $\Delta Volume$ and $\Delta Information$ do not necessarily correlate. In those cases, one of the speakers talks more but the dialogue topic is determined by the other speaker. This is the case with the MSDialog and MANTIS datasets, which contain information-seeking dialogues extracted from forums. More text is written by the Assistant but the repeated tokens are introduced by the Seeker. In the QuAC, SCS and WoW datasets, a different pattern is observed: the Assistant talks more and influences the conversation topic.

Repetition – who follows up on the topic introduced by another participant? In the majority of dialogue datasets across all dialogue types, the Assistant tends to lead in the number of repetitions (see the points above the y-axis in Figure 11). However, this is mostly characteristic of the information-seeking and task-oriented dialogues.

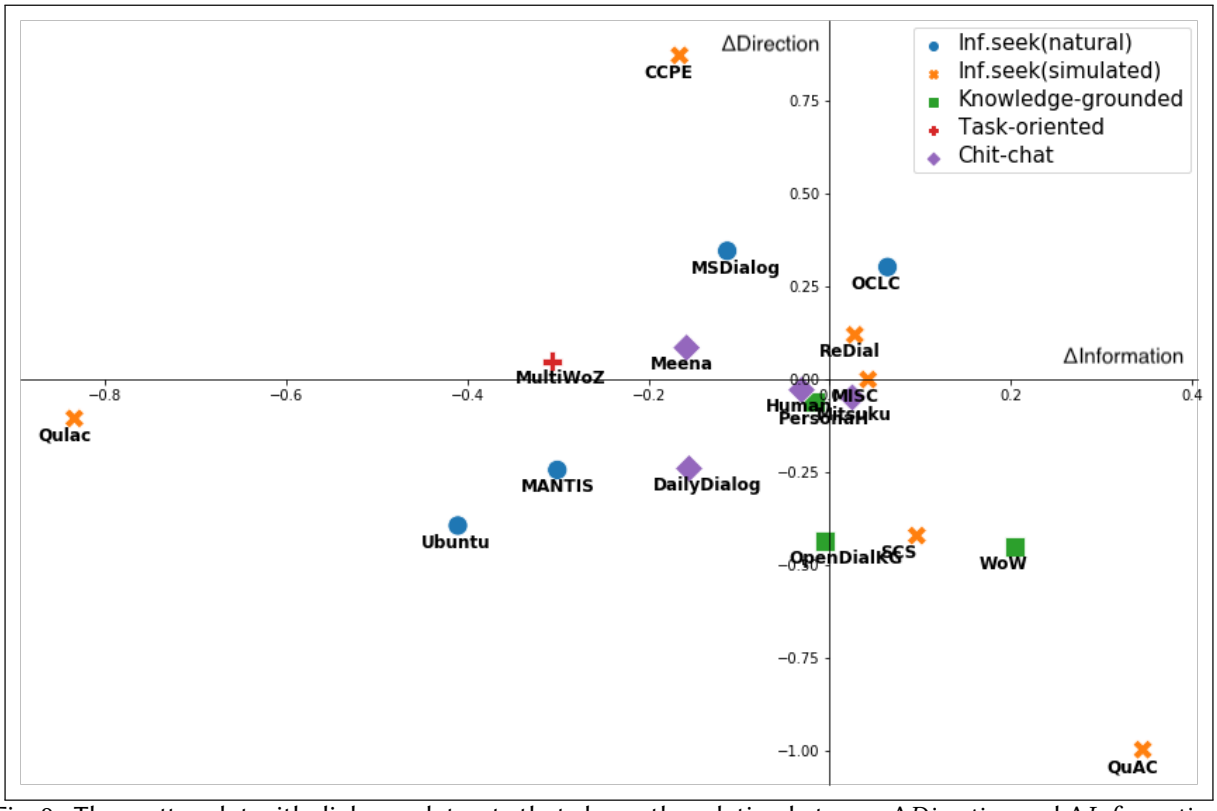


Fig. 9. The scatter plot with dialogue datasets that shows the relation between $\Delta Direction$ and $\Delta Information$.

To be able to simultaneously compare the datasets along all the dimensions of mixed initiative, we also present the results of the asymmetry metrics summarised into a single table (see Table 5). All Support dialogues we identified in the dialogue flow analysis in Section 5.1 have a positive $\Delta Direction$, while they do not display a shared pattern for the other three metrics. All Sharing dialogues have values close to zero, which we noticed already when looking at the individual plots. The (single) task-oriented dialogue dataset, while having a $\Delta Direction$ value close to zero, differs from the Sharing dialogues on all other dimensions of initiative. Finally, Search dialogues apart from the negative $\Delta Direction$ also predominantly have positive $\Delta Repetition$ (except for the WoW dataset).

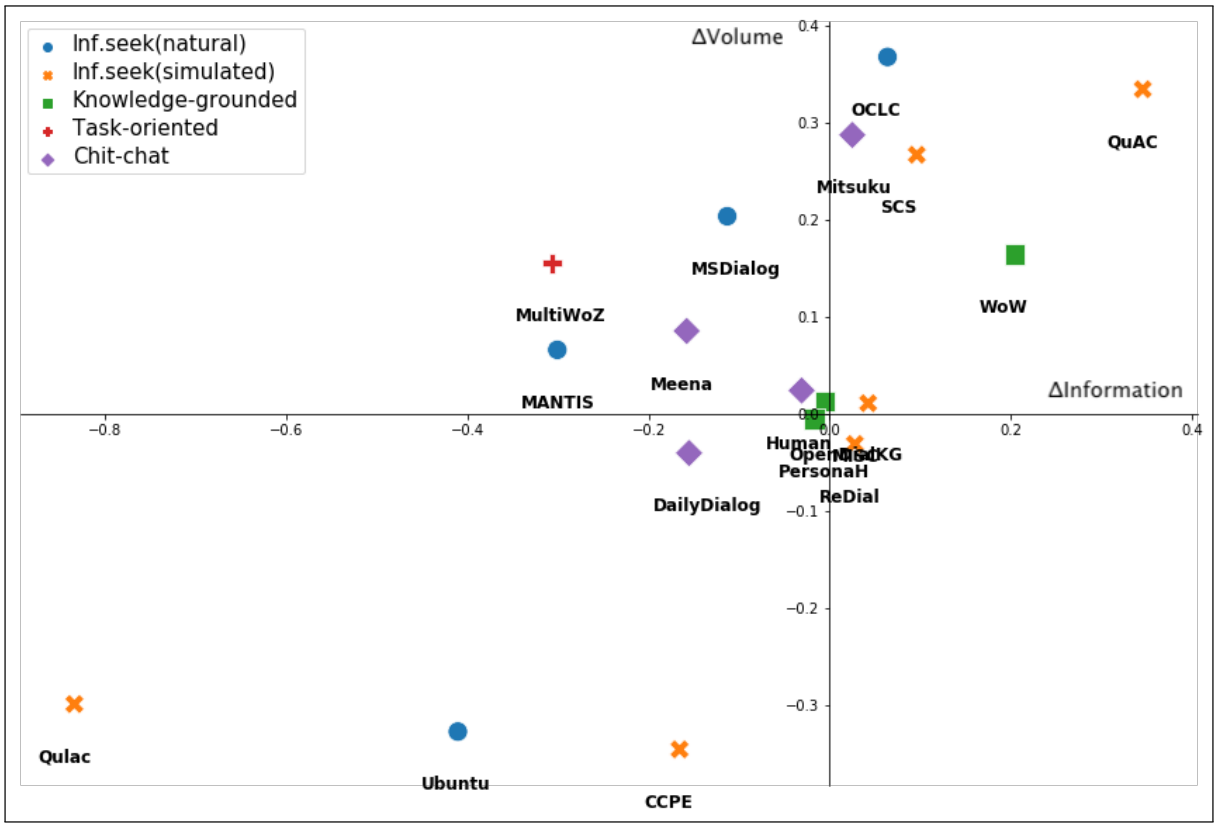


Fig. 10. The scatter plot with dialogue datasets that shows the relation between Δ*Volume* and Δ*Information*.

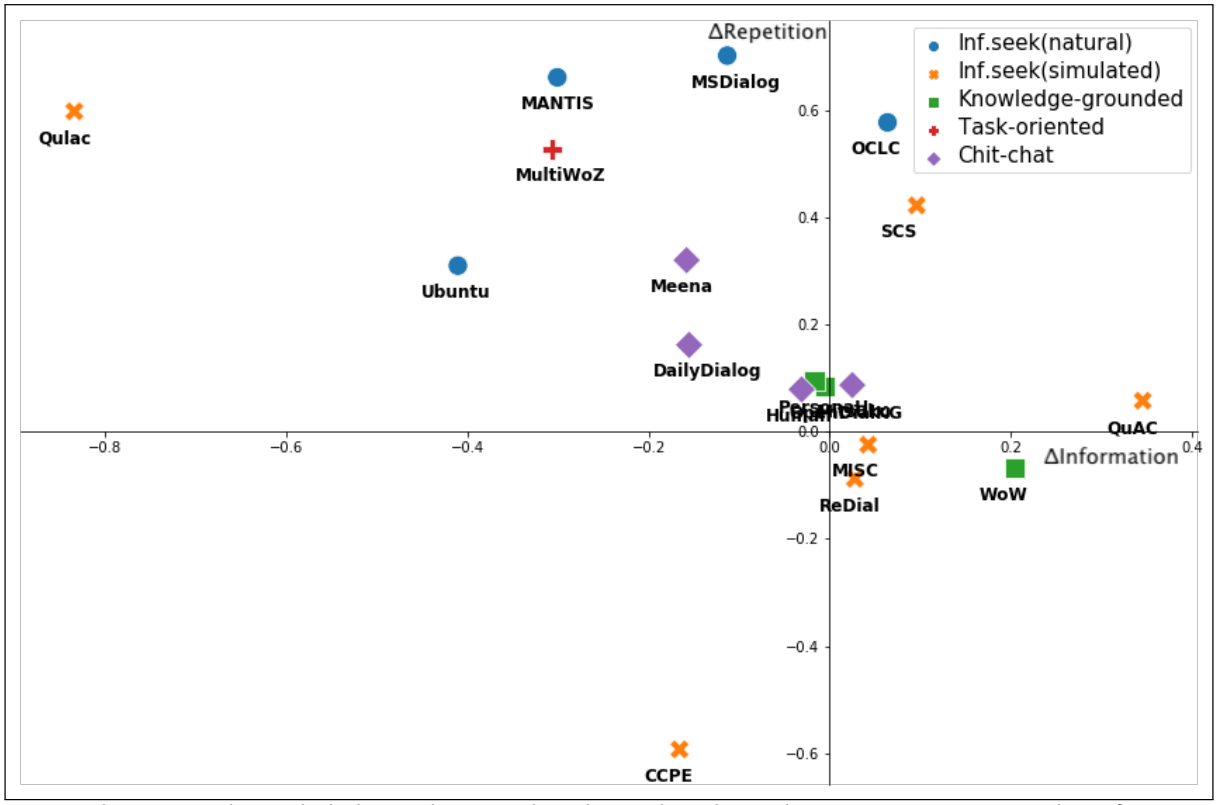


Fig. 11. The scatter plot with dialogue datasets that shows the relation between Δ*Repetition* and Δ*Information*.

It is clear from Table 5 that in Search dialogues the more questions come from the Seeker, i.e., $\Delta Direction \rightarrow -1$, the more information comes from the Assistant. This dependency is much less pronounced in the Support dialogues.

Note that in contrast to all other datasets, the OCLC dataset, which contains virtual reference interviews, appears at the top of the upper-right quadrant in all three scatter plots. The same result can be observed using Table 5: OCLC is the only dataset that has positive values across all four asymmetry metrics. It shows that professional intermediaries play a very active role on all dimensions of initiative: asking questions, providing information, and introducing new subtopics that engage the Seeker.

6 DISCUSSION

We started our analysis by introducing the dialogue datasets in Section 4.1. Our description of their sources and the purpose for which they were collected, gave rise to initial hypotheses about their content and the dialogue types they represent. We then proceeded to challenge these assumptions by systematically analysing and comparing structural patterns across dialogues. In the following, we summarise and discuss what we learned from this analysis with respect to (1) dialogue datasets for conversational search, (2) dialogue types and (3) relations between the dialogue analysis approaches we proposed.

6.1 Conversational Search Datasets

Obtaining samples that resemble the target interaction style of an information-seeking conversation is vital for the development and evaluation of conversational search systems. The results of our analysis show that the dialogues produced to inform the design of a conversational search system, such as SCSdata and MISC, do not reflect the interaction patterns observed in the transcripts of virtual reference interviews (OCLC). The differences observed are related to the distribution of mixed initiative between the dialogue participants. The Assistant is more passive in the simulated information-seeking dialogues, talking less and asking fewer questions (see Figures 5-11).

The results of our analysis indicate that the professional intermediaries in the OCLC dataset are more proactive than crowd workers and community experts, when measured on all dimensions of initiative. They lead the conversation not only by writing long responses ($\Delta Volume > 0$) and asking many follow-up questions ($\Delta Direction > 0$) but also by actively steering the topic of a conversation ($\Delta Information > 0$). These criteria should be considered in the evaluation of conversational search systems as well.

Our observations highlight that (1) the setup of the data collection tasks is crucial for obtaining representative dialogues; and (2) data analysis is crucial for verifying that the dialogues are representative. For example, the MISC dataset was designed to inform the conversational search task. However, we identified MISC as an information-sharing dialogue due to the symmetry of the speaker roles.

This finding is also in line with the observations reported by Trippas et al. [43], who concluded that MISC contains chit-chat and negotiations by analysing these dialogues using manual annotations. They suggest that the reason for this is the task setup. The participants were not explicitly instructed on how they should interact. The human intermediaries were provided with access to the information source but they were not instructed on how to conduct the interview. In contrast, professional librarians and other domain experts, such as call-center personnel, receive specialised training on how to efficiently identify relevant aspects of an information need, assist and guide the Seeker during the interview.

Our empirical evaluation demonstrates that the simulated information-seeking dialogue datasets, which were examined in our analysis, are not adequate for studying patterns of mixed initiative in

dialogue. We also showed that it is possible to perform the analysis of mixed initiative in dialogues automatically by comparing interaction patterns across dialogues. This allowed us to scale such an analysis to thousands of dialogue transcripts sourced from a dozen of publicly available datasets. The results show that neither of the publicly available datasets, which we considered in our analysis, matches the patterns extracted from the OCLC transcripts (see OCLC as an outlier in the scatter plots in Figures 9-11).

To move forward, the community should ensure that the conversational search systems that we design are modeled after professional intermediaries rather than the ad hoc strategies of non-expert volunteers. As empirically shown in our analysis, the communication strategies of expert librarians differ considerably from the ones employed by the intermediaries in the simulated scenarios.

It is important to note also that MISC is the only dataset in our analysis that contains dialogue transcripts produced by an automated speech recogniser. This likely lead to errors propagating from the utterance segmentation and classification steps into the analysis results. For example, we observed that MISC contains very few questions, also in comparison to other social datasets in Figure 8. Nevertheless, our findings were confirmed by the results of the manual analysis performed independently by Trippas et al. [43]. However, it is clear that to leverage data from spoken conversations our analysis techniques should be further adapted. Currently, there is a lack of spoken information-seeking dialogues annotated with utterance labels, which are required for training and evaluation of utterance classifiers.

6.2 Dialogue Types

Careful collection and annotation of training data is crucial for developing successful machine learning models [32]. Therefore, it is also of a great importance to systematically analyse and correctly classify dialogues prior to using them for model training and evaluation. This will allow us to avoid propagating biases towards undesirable patterns of interaction and dialogue characteristics in purely data-driven approaches, such as end-to-end dialogue models.

We uncovered important structural differences between the dialogue datasets that are generally considered to be of the same type. For example, the information-seeking dialogues that were collected in different settings may have very different structure of mixed initiative (see Figure 5).

Our results highlight that a common belief about a characteristic of a dialogue dataset can be incorrect. For example, DailyDialog is a popular dataset for training data-driven dialogue models and considered to contain chit-chat dialogues [5, 36].[2] These dialogues were collected from books for English learners and contain dialogues frequent in everyday situations, such as shopping or a job interview. In our analysis, we discovered that this dataset differs from other chit-chat datasets. This led us to discover that DailyDialog contains information-seeking dialogues, which also can be used to inform the conversational search task. Those results emphasise that the content of dialogue datasets is often poorly understood.

Such misconceptions about dialogue types may lead to wrong conclusions and misinterpretations of the experimental results. For example, Sinha et al. [36] assumed that DailyDialog is a chit-chat dataset, which led them to conclude that their machine learning model "captures the commonalities of chit-chat style dialogue" without investigating what the commonalities of chit-chat style dialogues really are.

Our results also indicate that information-seeking dialogues have similar dialogue flows as in task-oriented, grounded and chit-chat dialogues (see Figures 6-8). In most cases, apart from chit-chat, the initiative is skewed towards one of the dialogue participants.

[2]https://parl.ai/docs/tasks.html

Overall, our results confirm that the current approach to classifying dialogues (see Table 3) based on the task and the system architecture does not adequately reflect the dialogue properties, in particular the patterns of mixed initiative that to a large extent characterise the type of a dialogue interaction (see Table 5). Therefore, on the basis of our analysis, we introduced a new dialogue typology which reflects the communication strategies that are common across different dialogue datasets by measuring the degree of mixed initiative: Search, Sharing and Support (see Figures 6–8). While this classification schema is relatively simple, we show that it can be further enhanced by considering additional dimensions of initiative (see Figures 9–11, Table 5.)

6.3 Comparison of the Dialogue Analysis Approaches

The approaches to dialogue analysis we employed here shed light on the differences and similarities in patterns of mixed initiative. Dialogue flow diagrams provide a convenient overview of the transition frequencies between dialogue turns of different type. Their main benefit is that they provide a compact but explainable summary of the mixed initiative distribution. The important drawbacks of the dialogue flow diagrams are that they are difficult to compare and only reflect a single dimension of mixed initiative, namely, the utterance type. We overcome these limitations by introducing asymmetry metrics. In addition to quantifying the difference in utterance types, asymmetry metrics also reflect other dimensions of initiative: the distribution of utterance lengths and term repetitions.

The asymmetry metrics provide for sensible and meaningful dimensions that allow us to represent each dataset as a point in a vector space. For example, the scatter plot in Figure 9 was produced by using $\Delta Information$ and $\Delta Direction$ as x- and y-coordinates for each of the datasets. This approach allow us to conveniently compare datasets in a vector space and identify similar datasets using standard metrics, such as euclidean or cosine distance. This embedding approach also allows for results to be interpretable since we use the asymmetry metrics as dimensions, which reflect certain dialogue properties that are explicitly measured.

Importantly, the $\Delta Direction$ metric agrees with the results of our dialogue flow analysis. This is evident from the plot in Figure 9, where Search and Support dialogues are located on opposite sides of the x-axis. More specifically, all the datasets presented in Figure 7, i.e., MSDialog, OCLC, ReDial and MultiWoZ, have $\Delta Direction > 0$, while all datasets in Figure 6, i.e., SCSdata, MANTIS, Ubuntu, WoW, OpenDialKG and DailyDialog, have $\Delta Direction < 0$.

Both dialogue flow diagrams and asymmetry metrics can be used in a combination to successfully leverage their advantages and alleviate their drawbacks. For example, when designing a repository for dialogue data the asymmetry metrics can provide facets to enable search and browsing interface, while the dialogue flow diagrams can serve as a summary of the dataset content.

7 CONCLUSION

We introduced a dialogue representation approach and two dialogue analysis approaches, that is, dialogue flow diagrams and asymmetry metrics, that allow one to compare the structure of dialogues from different datasets. We applied both approaches to conduct a large-scale analysis of dialogue transcripts specifically focusing on the patterns of mixed initiative to distinguish information-seeking dialogues from other dialogue types. The results of our analysis including the dialogue fingerprints[3] as well as the scripts required to reproduce them[4] are publicly available to encourage future work in this direction.

[3]https://uvaauas.figshare.com/articles/dataset/dialogue_fingerprints/13356350
[4]https://github.com/svakulenk0/conversation_shape

7.1 Lessons Learned

Based on our analysis of the results and their discussion in the previous sections, we come back to answer the main research questions that were introduced in Section 1.

RQ1 What are the structural properties of information-seeking dialogues that differentiate them from other dialogue types?

Interestingly, we did not observe structural differences between information-seeking dialogues collected to inform the conversational search task and dialogues for other tasks, such as task-oriented, knowledge-grounded and chit-chat dialogues. On the contrary, we found that information-seeking dialogues vary a lot. We discovered several types of information-seeking dialogues and showed that they bear structural similarities to other dialogues.

On the basis of our analysis, we introduced a new dialogue typology that reflects the communication strategies that are common across different dialogue datasets: Search, Sharing and Support. We believe that conversational search systems should move from Search towards (information) Sharing and Support by taking over more initiative in a conversation, as exhibited in virtual reference interviews with professional librarians.

RQ2 Which datasets contain dialogues similar to virtual reference interviews with a professional librarian?

None of the dialogue datasets mirrors the patterns of mixed initiative in virtual reference interviews from the OCLC dataset. Thereby, we showed that the dialogues that were collected in a laboratory environment between volunteers using simulated information needs do not represent the dialogue type that naturally occurs between an expert intermediary and a seeker with a genuine information need. Non-expert intermediaries write less and ask less questions than professional librarians. This finding implies that the data collection procedures that the community has designed to inform the conversational search task requires adjustment and better quality control.

Thus, the results of our analysis provide evidence for the claim that existing datasets collected to inform the conversational search task (MISC, SCSdata, etc.) are not suitable for studying and designing mixed initiative systems, as we argued in Section 6.1. The community should focus on more realistic datasets, such as OCLC, to better understand the patterns of initiative from interactions between skilled interviewers/librarians and information seekers.

Overall, we showed that dialogue flow diagrams can provide a convenient overview for a dialogue dataset and they are easy to interpret, while asymmetry metrics allow us to conveniently compare dialogue datasets across several dimensions of mixed initiative simultaneously. Both approaches enabled us to discover frequent patterns that provide valuable insights as to the important characteristics and qualities of the dialogue datasets currently available to the research community. We consider this to constitute an important milestone towards establishing a mechanism for continuous refinement of our understanding of information-seeking dialogue interactions based on the growing volume of empirical data, while also evaluating the quality of the dialogue data being used for model development.

7.2 Future Work

The results of our large-scale dialogue analysis are based on automatic utterance classification and term-based repetition patterns. Both of these approaches have their limitations and the errors introduced by them affect the results of dialogue analysis. The metrics we introduced for measuring initiative in dialogue are relatively simple and crude. We are likely to underestimate the number of questions by both participants due to errors in classification and the number of repetitions since the lexical matches capture only explicit references to the previous context overlooking more subtle semantic relations between the utterances. Nevertheless, our results demonstrate

the value of automatic data analysis approaches that have implications on the data collection, interaction design and evaluation. Future work should focus on the evaluation and improvement of the representational power of the dialogue analysis approaches.

The interaction patterns we discovered can also be used to enhance browsing and search interfaces providing access to a repository of dialogue datasets. Modeling structural properties of dialogues based on their content rather than merely their metadata description will help to retrieve samples that can be used for training and evaluation of dialogue models.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for extensive comments and suggestions that helped us to improve the paper.

This research was inspired by the discussions held at the Dagstuhl Seminar 19461 on Conversational Search. We thank the organisers and participants of the seminar, and, especially, Filip Radlinski and Nicholas Belkin for triggering the idea for this study. We are especially grateful to Marie Radford, Lynn S. Connaway and Andrew K. Pace for providing us with the dataset of virtual reference interviews from OCLC.

This research was supported by the Netherlands Organisation for Scientific Research (NWO), Innovational Research Incentives Scheme Vidi (016.Vidi.189.039) and Smart Culture - Big Data/Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the Google Faculty Research Awards program, and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. 475–484.

[3] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational Search (Dagstuhl Seminar 19461). *Dagstuhl Reports* 9, 11 (2020), 34–83.

[4] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing Agent-human Interactions during the Conversational Search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval*.

[5] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2019. Plato: Pre-trained dialogue generation model with discrete latent variable. *arXiv preprint arXiv:1910.07931* (2019).

[6] Nicholas J Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. 1995. Cases, Scripts, and Information-seeking Strategies: On the Design of Interactive Information Retrieval Systems. *Expert systems with applications* 9, 3 (1995), 379–395.

[7] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation* 38, 2 (1982), 61–71.

[8] Frans AJ Birrer. 1998. Asymmetry in the Dialogue Between Expert and Non-Expert. *Proceedings of the International Society for the Study of Argumentation* (1998).

[9] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. 5016–5026.

[10] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, Brussels, Belgium, October 31 - November 4, 2018. 2174–2184.

[11] Jennifer Chu-Carroll and Michael K. Brown. 1998. An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions. *User Modeling and User-Adapted Interaction* 8, 3-4 (1998), 215–254.

[12] Robin Cohen, Coralee Allaby, Christian Cumbaa, Mark Fitzgerald, Kinson Ho, Bowen Hui, Celine Latulipe, Fletcher Lu, Nancy Moussa, David Pooley, Alex Qian, and Saheem Siddiqi. 1998. What is Initiative? *User Modeling and User-Adapted Interaction* 8, 3 (1998), 171–214.

[13] Penny J. Daniels, Helen M. Brooks, and Nicholas J. Belkin. 1985. Using Problem Structures for Driving Human-computer Dialogues. In *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 1985, 1st International Conference, University of Grenoble, France, March 18-20, 1985. Proceedings*. 645–660.

[14] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020. The Second Conversational Intelligence Challenge (ConvAI2). In *The NeurIPS'18 Competition*. Springer, 187–208.

[15] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

[16] Eric N. Forsythand and Craig H. Martell. 2007. Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007), September 17-19, 2007, Irvine, California, USA*. 19–26.

[17] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *arXiv preprint arXiv:2101.09459* (January 2021). https://arxiv.org/abs/2101.09459

[18] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural Approaches to Conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (2019), 127–298.

[19] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. *arXiv preprint arXiv:2004.00646* (2020).

[20] Geoffrey Leech and Martin Weisser. 2013. The SPAADIA Annotation Scheme. Available from http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf.

[21] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. 9748–9758.

[22] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*. 986–995.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[24] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*. 285–294.

[25] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 845–854.

[26] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: A Novel Multi-Domain Information Seeking Dialogues Dataset. *arXiv preprint arXiv:1912.04639* (2019).

[27] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. 989–992.

[28] Marie L Radford and Lynn Silipigni Connaway. 2013. Not Dead Yet! A Longitudinal Study of Query Type and Ready Reference Accuracy in Live Chat and IM Reference. *Library & information science research* 35, 1 (2013), 2–13.

[29] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*. 353–360.

[30] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*. 117–126.

[31] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.
[32] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* (2019).
[33] Tefko Saracevic, Amanda Spink, and Mei-Mei Wu. 1997. Users and Intermediaries in Information Retrieval: What are They Talking About?. In *User Modeling*. Springer, 43–54.
[34] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. In *NAACL*.
[35] Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. (2020), 2430–2441.
[36] Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an Unreferenced Metric for Online Dialogue Evaluation. *ACL* (2020). https://arxiv.org/abs/2005.00583
[37] Stefan Sitter and Adelheit Stein. 1992. Modelling the Illocutionary Aspects of Information-Seeking Dialogues. *Information processing & management* 28, 2 (1992), 165–180.
[38] Adelheit Stein, Jon Atle Gulla, and Ulrich Thiel. 1999. User-Tailored Planning of Mixed Initiative Information-Seeking Dialogues. *User Modeling and User-Adapted Interaction* 9, 1-2 (1999), 133–166.
[39] Robert S. Taylor. 1968. Question-Negotiation and Information Seeking in Libraries. *College & Research Libraries* 29, 3 (1968).
[40] Paul Thomas, Mary Czerwinski, Daniel McDuff, and Nick Craswell. 2020. Theories of Conversation for Conversational IR. In *SIGIR 3th International Workshop on Conversational Approaches to Information Retrieval (CAIR'20)*.
[41] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A Data Set of Information-seeking Conversations. In *Proceedings of the 1st International Workshop on Conversational Approaches to Information Retrieval*.
[42] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*. 32–41.
[43] Johanne R Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2020), 102162.
[44] Svitlana Vakulenko. 2019. *Knowledge-based Conversational Search*. Ph.D. Dissertation. Faculty of Informatics, Technische Universität Wien.
[45] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2020. An Analysis of Mixed Initiative and Collaboration in Information-Seeking Dialogues. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. 2085–2088.
[46] Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A Data-Driven Model of Information-Seeking Dialogues. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*. 541–557.
[47] Marilyn A. Walker and Steve Whittaker. 1990. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In *28th Annual Meeting of the Association for Computational Linguistics, 6-9 June 1990, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, Proceedings*. 70–78.
[48] Steve Whittaker and Phil Stenton. 1988. Cues and Control in Expert-Client Dialogues. In *26th Annual Meeting of the Association for Computational Linguistics, 7-10 June 1988, State Univerity of New York at Buffalo, Buffalo, New York, USA, Proceedings*. 123–130.
[49] Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. 418–428.

7. Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation

Bibliographic Information

Svitlana Vakulenko, Bill Byrne, Adrià de Gispert: Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation. ECAI 2023: 2378-2385.

Uniform Training and Marginal Decoding for Multi-Reference Question-Answer Generation

Svitlana Vakulenko[a;*], **Bill Byrne**[ab] **and Adrià de Gispert**[a]

[a]Amazon Alexa AI
[b]University of Cambridge

Abstract. Question generation is an important task that helps to improve question answering performance and augment search interfaces with possible suggested questions. While multiple approaches have been proposed for this task, none addresses the goal of generating a diverse set of questions given the same input context. The main reason for this is the lack of multi-reference datasets for training such models. We propose to bridge this gap by seeding a baseline question generation model with named entities as candidate answers. This allows us to automatically synthesize an unlimited number of question-answer pairs. We then propose an approach designed to leverage such multi-reference annotations, and demonstrate its advantages over the standard training and decoding strategies used in question generation. An experimental evaluation on synthetic, as well as manually annotated data shows that our approach can be used in creating a single generative model that produces a diverse set of question-answer pairs per input sentence.

1 Introduction

A significant limitation of current question generation models is that they are trained to produce only a single question based on a given input context [19, 13]. This is a shortcoming in that many sentences, particularly in the news or science domains, are long and information-dense. These sentences can easily give rise to multiple interesting and useful questions. While standard question generation methods can produce multiple questions, these are usually paraphrases of the same question rather than semantically distinct questions.

Improved diversity of generated questions was identified as an important research direction in a literature review covering the recent research on question-answer generation (QAG) [19]. The authors note that the standard approaches primarily focus on generating one question as a 1-to-1 mapping problem and consequently, such methods fail to generate multiple diverse questions. The major challenge they see with respect to diverse QAG is in *identifying different question-worthy context words dynamically* and actively controlling the question generation process.

Automatically generated questions are useful for fine-tuning question answering (QA) models [17, 14] and augmenting document collections to improve matching [10, 1, 16, 3]. In supervised learning, richer training sets are likely to lead to more robust and diverse QA models. Generating a diverse set of questions was also shown to improve performance on non-QA downstream tasks that require general passage representations, such as paraphrase detection, named entity recognition and sentiment analysis [7].

To enable diverse QAG, we posit the task of generating a set of questions from an input context, such as a single sentence or a paragraph of text. We are interested in generating a diverse set of questions that focus on different aspects of the input text rather than mere paraphrases of the same question. To achieve this, we make use of the simple observation that different questions tend to have different answers. Using different answer spans extracted from the input text can ensure generation of a diverse set of questions.

To the best of our knowledge, we are the first to propose an approach for extending a state-of-the-art QAG model to a multi-reference setting, i.e., a training scenario in which multiple correct output sequences are available for matching the same input sequence. We also streamlined training and evaluation of our approach by generating a multi-reference QA dataset centered around named entities.

Named entities, such as names, locations and dates, are suitable targets for fact-based QA, and related work on question generation often generates questions with answers that are named entities [4, 18]. When there are multiple entities per sentence, our approach can generate automatically a dataset with multiple distinct QA pairs. We will use a combination of an off-the-shelf named entity recognition (NER) model and a standard QAG model to produce synthetic datasets for our experiments.

The goal of our work is to develop variations in model training and decoding to account for multiple QA pairs in the ground truth. The NER model here is used as a proxy for generating a multi-reference dataset as a proof-of-concept for our approach. More question-answer pairs can be obtained by conducting manual annotations or using other models, such as keyword extraction and summarization. Even simple heuristics such as noun phrase extractors can help to identify answer spans that potentially lead to useful questions.

While most of the related work focuses on generating questions given a passage [10, 12], we take a more ambitious goal of generating multiple diverse QA pairs given a single sentence. It is more difficult since the search space is much smaller and the likelihood of generating a duplicate is much higher. None of the existing QAG datasets, such as SQuAD [12] or Natural Questions [8], has an exhaustive list of all possible questions per sentence, except for Monserrate [13], which we use in our evaluation.

Our evaluation is also the first one to focus on all of the following aspects: diversity, correctness and completeness of the generated QA set, the later being traditionally overlooked property of the generative models. Using synthethic data generated with NER allowed

* Corresponding Author. Email: svvakul@amazon.com.

us to evaluate recall of the reference QA set. The results show that our approach is able to strike a fine balance between completeness and diversity, i.e., able to match the reference set without generating duplicate questions. Such cost efficiency is vital for auto-regressive models aiming to minimise the number of redundant token generations.

Our contributions in this paper are three-fold:

- an automated approach for generating a multi-reference corpora of QA pairs using a pre-trained NER model;
- a new public dataset generated using the proposed approach as well as the code required to reproduce and extend it;
- modifications to the standard training and decoding procedures that allow us to train an end-to-end QAG model able to generate a set of diverse QA pairs as confirmed by the results of our comprehensive experimental evaluation[1].

2 Background

There are two types of question generation tasks: answer-aware and answer-agnostic [19]. In the first case, the answer is given defining the focus of the generated question. The latter case is more realistic since the answers are not given in practice and the model should be able to generate both questions and answers.

2.1 Question-answer generation

Given the context C as input (such as a sentence or a paragraph of text), the goal of the QAG model f_{QAG} is to generate a question-answer pair as an output:

$$f_{QAG}(C) =< Q, A > \tag{1}$$

The state-of-the-art approaches to question generation consider this as a sequence-to-sequence task. For our setup, we choose a sequence-to-sequence model that is trained end-to-end to generate the QA pairs jointly, as a single demarcated output sequence. It has furthermore been shown advantageous to generate a QA pair in reverse order by first producing the answer given the context and then producing the question given both the context and the answer embeddings just produced by the model [5]. This approach effectively makes question generation answer-aware and conditioned on the latent answer representation so as to account for variance in prediction and to help train both answer and question generation at the same time.

2.2 Decoding

A standard sequence-to-sequence model uses autoregressive generation. At every step in decoding, the model produces a probability distribution over the entire vocabulary, i.e., the probability for each of the possible tokens to be generated given the input and the output sequence that was already generated so far. For every token at position i, the model produces a distribution such that for the j-th token in the vocabulary $v_j \in V$ there is a probability assigned to its appearance at this position as p_{ij}, where

$$p_{ij} = P(t_i = v_j | t_0 ... t_{i-1}) \tag{2}$$

[1] The annotation guidelines, datasets and code for our experiments are available at: https://github.com/amazon-science/multi-reference-qa-generation.

The simplest decoding approach is called greedy because it always picks the token with the highest probability at each position i using `argmax`. As an alternative, sampling approaches, which make use of the entire probability distribution, have been shown to outperform greedy decoding in practice [2, 6]. To avoid sampling low-probability tokens from the tail of the distribution, the distribution is usually truncated, typically to the top-k tokens as ranked by their probabilities. This decoding approach is called *top-k sampling* and is often applied in practice [2]. The tokens are sampled from the truncated distribution:

$$\begin{gathered} V_i^{topk} = \{v_j \in V : |\{j' : p_{i,j'} > p_{i,j}\}| \leq k\} \\ \widetilde{p}_{i,j} = \frac{p_{i,j}}{\sum_{j' \in V_i^{topk}} p_{i,j'}} \text{ for } j \in V_i^{topk} \\ t_i \sim \widetilde{p}_{i,j} \end{gathered} \tag{3}$$

where $V_i^{topk} \subset V$ are the k most probable tokens predicted by the model for position i, and t_i is the token produced at position i by drawing from V_i^{topk} according to $\widetilde{p}_{i,j}$.

2.3 Training

The model is trained using the cross-entropy (CE) loss that assigns the highest probability to the exact next token as given in the ground-truth reference sequence:

$$CE_O = -\sum_{i=1}^{|O|} \sum_{j=1}^{|V|} t_{ij}^* \log p_{ij} \tag{4}$$

where t_i^* is a $|V|$-dimensional one-hot vector encoding the ground truth target token at position i, and $|O|$ and $|V|$ are the number of tokens in the output sequence and in the vocabulary, respectively.

3 Approach

Our approach consists of the three main parts:

(1) an *augmentation* procedure that generates a dataset with multiple distinct QA reference pairs per input sequence;

(2) a *training* procedure that makes use of those reference pairs by modifying the target output distribution; and

(3) a *decoding* procedure that allows us to produce multiple distinct QA pairs per input sequence using this output distribution.

3.1 Multi-reference augmentation

The first step of our approach is generating a dataset that we use for training and evaluation of the QAG models. We employ a pipeline approach to achieve this by using the output of a separate answer extraction model as input to our pre-trained QAG model.

First, the answer extraction model f_{AE} produces a set of answer spans derived from the input context:

$$f_{AE}(C) = \{A_0, ..., A_l, ..., A_n\}, \text{where } A_l \subset C \tag{5}$$

In our experiments, we use an off-the-shelf NER model as our answer extractor from sentences but it could be another model, such as a summarization model, which identifies salient text spans.

Secondly, for each of the extracted answers we proceed to generate the rest of the output sequence, i.e., the corresponding question:

$$\forall A_l \in f_{AE}(C) : Q_l = f_{QAG}(< C, A_l >) \tag{6}$$

It is possible to reuse the same model trained on the original QAG task defined in (1) for this task as well by simply using each of the answers A_l separately as an input to the decoder of the QAG model. For example, given an input sentence from a news article: *"Comcast rolled out a Web-based on-demand television and movie service on Tuesday that gives customers access to more than 2,000 hours of television and movies."*, we first use NER to detect all named entities in the sentence: "Comcast", "Tuesday" and "more than 2,000 hours". Then, we use each of those named entities as input to the decoder of a pre-trained model to generate the rest of the output sequence as the corresponding question. We generate a single question for each of the answer-entities, resulting in three QA pairs for this example, such as:

f_{QAG} *("Comcast rolled out a Web-based on-demand television and movie service on Tuesday that gives customers access to more than 2,000 hours of television and movies."[BOS]"**a**> Comcast **q**> ") = "Who rolled out a Web-based on-demand television and movie service?"*

Thereby, we obtain a *set* of n QA pairs and recast the original QAG task defined in (1) into the updated QAG task f_{QAG} where the output target is a set of QA pairs instead of a single QA pair:

$$f_{QAG}(C) = \{< Q_l, A_l >\}_{l=[1;n]} \tag{7}$$

For a sequence-to-sequence QAG model we described in Section 2.1, one pair is produced by generating a text sequence starting from the answer and followed by the question with a separator in-between. It means that given a single input sequence C the goal of the model is now to produce a *set* of n output sequences instead:

$$f_{QAG^*}(C) = \{A_l \cdot Q_l\}_{l=[1;n]} \tag{8}$$

where $\cdot$ indicates concatenation.

3.2 Uniform training

Similar to the standard sequence-to-sequence training, we use the target question-answer pairs for training. Our proposed modification accounts for multiple correct output sequences in the case of question generation making use of the positional information of the answer in the beginning of the output sequence.

To train the model on n QA reference pairs, we propose to modify the target output distribution for the first token in the output sequence t_0^*, which is used in the CE loss, see Equation 4 in Section 2.3. More specifically, given a training sample $(C, \{A_l \cdot Q_l\}_{l=[1;n]})$ we construct a subset of tokens:

$$T_A = \cup_{l=1}^{n} \{A_l[0]\} \tag{9}$$

i.e., the set consisting of the first tokens of every answer sequence.

In uniform training, the target distribution in Equation 4 is modified so that in position 0 it becomes

$$t^*_{0,j} = \begin{cases} \frac{1}{|T_A|} & v_j \in T_A \\ 0 & v_j \notin T_A \end{cases} \tag{10}$$

Hence, instead of training the model to produce a single token with probability 1, we are training it to produce *all* tokens in T_A with the uniform probability distribution among them and with 0 probability assigned to the tokens outside of this set (see Figure 1 (a). Thereby, we communicate to the model that all the answers are equally likely to begin at this step.

For the example that was introduced in Section 3.1, the model will be trained to predict the first tokens of the answers: "_Com", "_Tuesday" and "_more" given the sentence as input to the encoder and prefix "a> " as input to the decoder.

We use uniform training only on the set T_A, which is constructed using the first tokens of the output sequences. The rest of the sequences are generated using the standard teacher forcing approach. We implemented this training procedure via multi-task learning, in which different training batches are mixed in equal proportions. Importantly, T_A is used only for training, at inference time the method is agnostic to the answer options and generates a set of QA pairs in parallel.

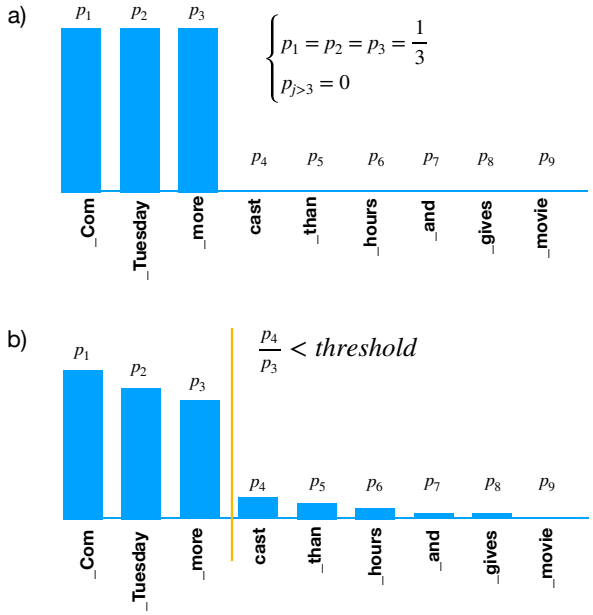


Figure 1: Modified training and decoding procedures for generation of multiple question answer pairs: (a) uniform training, and (b) marginal decoding.

3.3 Marginal decoding

To better leverage the training procedure described in the previous section, we devise a matched decoding procedure that builds on the expectation that, at the first step of decoding, the model will produce a probability distribution with the shape described in Equation 10. More specifically, we are looking for the top n tokens predicted for t_0, the token in the first position of the output sequence:

$$T'_A = V_0^{topk} \tag{11}$$

following Equation 3.

After the set T'_A is produced, we then proceed to generate a set of n sequences independently each starting with a different token from the set: $t_{0,j} \in T'_A$. We can use any of the standard decoding approaches available, such as greedy decoding or beam search, to decode the rest of each sequence.

Since the number of such tokens, i.e., the number of QA pairs, is unknown at inference time, n can be chosen as a hyper-parameter. This approach makes the output similar to other decoding approaches that produce multiple sequences by sampling from the distribution. However, a more sensible approach is to make use of the distribution

P itself to determine the cut-off threshold by looking at the sorted token probabilities themselves (see Figure 1 (b).

The probability threshold could be set as an absolute value, e.g., stop considering answer tokens if the difference between successive probabilities is more than a threshold value. However, since the probabilities assigned to the correct tokens vary in our case, depending on the size of the correct answer set (as specified per Equation 10), it is more appropriate to compare probability ratios instead. A snippet with the pseudo-code for our decoding algorithm with a relative probability threshold is given in Algoritm 1.

Algorithm 1: Marginal decoding

```
Input: C, threshold ;
;
// produce and sort the probability distribution for t_0 ;
P ← f_QAG*(C) ;
P ← sort(P) ;
;
// add the most probable token for t_0 to a set ;
T'_A ← { } ;
(t, lastp) ← P.pop() ;
T'_A.add(t) ;
;
// keep iterating over the sorted probability distribution
// adding the most probable tokens until the threshold is
   reached
for (t, p) in P do
    if p / lastp < threshold then
        break ;
    T'_A.add(t) ;
    lastp ← p ;
;
// produce the rest of the output sequences using a standard
   decoding approach ;
outputs ← { } ;
for t in T'_A do
    s ← C + "[BOS]" + t ;
    outputs.add(decode(f_QAG*, s))
```

The relative threshold changes the number of QA pairs produced by the model based on the output distribution of the first answer token. The more samples we produce the more likely we would fit the ground truth samples, i.e., recall increases but we are also more likely to generate duplicates, i.e., diversity decreases. We experimented with different thresholds but set it in our experiments such that our model produces on average the same number of QA pairs as the baseline approaches to ensure a fair comparison.

4 Automated Evaluation

Our evaluation results are designed to reflect the ability of the QAG model to generate a complete set of question-answer pairs given an input sentence. For every QA pair in the ground truth, we find the most similar QA pair produced by the QAG model.

Note that we had to modify ROUGE calculation for this task to work with the multi-reference dataset. It measures correctness for each of the QA sample from the ground truth separately and hence also indicates recall of the whole set (see Section 4.4 for the description of the metric). We manually examine the generated samples that differ from ground truth in more detail in Section 5.

4.1 Datasets

We use two datasets for evaluation: a large generated dataset, NewsQA-Entities, and a small manually annotated, Monserrate.

NewsQA-Entities is the benchmark we generated using the approach proposed in Section 3.1. The original NewsQA dataset is not a multi-reference dataset. It contains a single question per sentence for the majority of questions. We augment it with synthetic question-answer pairs to simulate a multi-reference dataset that can be then used both for training and evaluation purposes.

For every sentence of the article that contain answers to questions from NewsQA, i.e., potentially newsworthy for the reader, we detect mentioned entities using a pre-trained NER model. Those entity mentions are then used as prefixes (short answers a>) for our QAG model to generate questions.

To produce this benchmark, we use:

- the test and train splits of NewsQA [15], which contains manually produced QA pairs for sentences from news articles;
- Stanza English for NER[2];
- T5 base QAG model fine-tuned on the train split of NewsQA with a greedy decoding strategy.

Stanza was chosen as a state-of-the-art NER model with F1 score of 92% on news articles [11]. In our approach, this model can be substituted with a summarizer, clause extractor or manual human annotations, whichever is deemed more appropriate to provide examples of desired answers for the given dataset.

We select only the sentences that contain at least one entity detected by Stanza and use a random subset of 11K such sentences for training the model and 5k sentences for testing. 41% sentences in the test split contain more than two named entities (see Table 1 for other summary statistics).

Monserrate is a corpus specifically built to evaluate question generation systems [13]. It contains sentences that originate from two English texts about Monserrate palace and sets of questions manually annotated for each of those sentences. The answers to those questions are not provided. The questions in Monserrate are designed to provide an exhaustive reference set for each of the input sentences. The authors use it to evaluate performance of the state-of-the-art systems but do not quantify how many of the reference questions those systems can actually recall.

Monserrate has only 73 sentences annotated and hence is not suitable for training the model. There are, on average, 26 questions per sentence, and 1,933 questions, in total. In contrast with our NewsQA-Entities datasets, the questions in Monserrate tend to overlap, i.e., many of the questions are paraphrases rather than semantically distinct questions. Some of the questions in Monserrate are unanswerable given the reference sentence. Therefore, we prepare an annotation task to filter out such questions and group paraphrases into distinct subsets.

The annotators resolved all the disagreements (Cohen's kappa of 0.39) by discussing the samples that were labeled differently. After the annotation was completed, we obtained a new version of the dataset with 9 semantically distinct questions per sentence, on average, and 647 questions, in total. We use all the questions including paraphrases for matching the generated questions but evaluate recall on the distinct questions only.

[2] https://stanfordnlp.github.io/stanza

Table 1: Statistics of the NewsQA-Entities dataset with the most common entity types. Abbreviations: GPE (Countries/cities/states), ORG (Organization), NORP (Nationalities/religious/political group).

Split	#Sentences	#QA pairs	PERSON	GPE	ORG	DATE	CARDINAL	NORP	TIME	ORDINAL
train	11,203	29,024	27%	17%	15%	14%	8%	6%	2%	2%
test	5,000	13,086	25%	16%	16%	15%	9%	6%	2%	2%

4.2 Models

We evaluate the following three models fine-tuned on the QAG task: 1) the model trained on the original NewsQA dataset; 2) the same model further fine-tuned on our multi-reference dataset using the standard approach; 3) the same model but fine-tuned using the proposed uniform training approach.

Trained on NewsQA. A T5-base model for autoregressive text generation trained on the original NewsQA dataset. We use sentences containing the answers as input and concatenated QA pairs as output of the model. This is also the same model that was used for data augmentation to produce the NewsQA-Entities dataset.

Fine-tuned on NewsQA-Entities. The same baseline model further fine-tuned on the generated NewsQA-Entities dataset.

Fine-tuned on NewsQA-Entities + Uniform. We observe that while fine-tuning the model using uniform distribution (Section 3.2) over the first answer tokens, makes the model forget the original task. Therefore, we train it in the multi-task scenario by mixing both types of batches in equal proportions.

All three models were trained with the maximum input length of 512 tokens and the maximum output length of 1,024 tokens. We selected these sequence lengths by calculating the dataset statistics of the NewsQA training split. The maximum number of epochs was set to two with early stopping if the loss converges before that.

4.3 Decoding strategies

We test the standard decoding strategies against our approach.

Greedy: always picks the argmax token and generates a single output sequence.

Beam search: top-k hypotheses with max product probability (we set number of beams to 5 in our experiments).

Top-k sampling: samples from the top-k tokens ordered by probability (we set k=50 in our experiments).

Top-p (nucleus) **sampling**: considers the top tokens for which the sum of their probabilities is at least p (in our experiments p=0.92).

Marginal decoding: our approach described in Section 3.3.

For marginal decoding we generate the same number of QA pairs as the maximum number of QA pairs per sentence based on the dataset statistics (7 for NewsQA-Entities and 21 for Monserrate) since it uses a cut-off threshold to determine the number of output sequences dynamically. For all other sampling methods, except for greedy search, we use the same number of output sequences as generated by the marginal decoding approach to make their results more comparable. Since the number of QA samples have a direct effect on the performance metrics, optimising for it is important for the end task but for the purpose of our evaluation it should be fixed.

4.4 Evaluation metrics

We are evaluating diversity of the generated questions and answers separately and also assessing how many of the ground-truth sample our generative model can recall.

Distinct-1 is the standard metric used for measuring diversity in NLG, e.g. for the dialogue response generation task [9]. It is defined as the number of distinct tokens (unigrams) divided by the total number of generated tokens. We calculate it by measuring the proportion of the unique tokens across all the sequences generated for the same input sequence. Greedy decoding always produces a single output sequence. Hence, its diversity metric is always 1.

ROUGE is the standard metric used in summarization. We calculate ROUGE-L F-measure that measures the token overlap for the longest common subsequence of a pair of strings. Since we have multiple different reference sequences, we take the maximum ROUGE score calculated across all the predicted sequences. Then, we compute the mean of these ROUGE scores across all the test samples. Note that we measure ROUGE with respect to each of the reference sequences to evaluate how well we can recall the ground-truth questions and answers on average. Hence, the number of scores being aggregated will always be the same irrespective of the number of questions generated by the model. We measure ROUGE only for questions with distinct answers to avoid matching the same question paraphrases multiple times.

softM0.5 shows a proportion of the generated sequences that have a certain overlap with the target sequence, which is defined by the minimum ROUGE score. We use ROUGE-L F-measure of 0.5 here to count the sequence as a partial match for both questions and answers. This is a novel metric we introduce to provide more insight into our results. Since ROUGE scores are aggregated across all references, it is impossible to tell whether some of the references were matched very well and others not at all, or all of the references were matched to some extent. Thereby this soft match score allows to estimate the percentage of references that we consider as a partial match determined by the similarity threshold.

SBERT measures semantic similarity between questions beyond the token overlap measured by ROUGE. We use a pre-trained model available from the SentenceTransformer library trained on the Quora Duplicate Questions dataset using DistilBERT base model.[3].

4.5 Results on NewsQA-Entities

Results of the automated evaluation are given in Table 2. The table demonstrates that both questions and answers generated using our approach are on average more diverse and match the ground truth better than all the other baselines (see last raw of this table).

Most of the answers produced by the original model trained on NewsQA are not entities. Often those are rather longer spans extracted from the input sentence (see Table 4).

Greedy decoding produces only a single output sequence and, therefore, receives the diversity score of 1 for answers and questions. While both top-p and top-k sampling lead to lexically diverse questions, their answers tend to overlap.

Marginal decoding in the combination with uniform training (last row of Table 2) has the highest diversity scores and fits the ground-truth data better (ROUGE-L mean), especially in terms of the answer spans. This means that our approach succeeds in emulating the NER-based generation pipeline and produces diverse questions with entities as their answers.

[3] https://www.sbert.net

Table 2: Evaluation results on NewsQA-Entities. #QA is the average number of QA pairs generated per input sentence.

Model	Decoding	#QA	Answers			Questions			
			Distinct-1	ROUGE	softM0.5	Distinct-1	ROUGE	softM0.5	SBERT
Ground truth		2.6	**0.99**	**1**	**1**	**0.78**	**1**	**1**	**1**
Trained on NewsQA	Greedy	**1**	**1**	0.25	0.24	1	0.43	0.33	0.61
	Beam search	3	0.42	0.25	0.19	0.51	0.37	0.27	0.58
	Top-k sampling	3	0.61	0.37	0.33	**0.75**	0.38	0.26	0.62
	Top-p sampling	3	0.6	0.36	0.33	0.74	0.38	0.25	0.61
	Marginal decoding	2.7	0.73	0.36	0.31	0.67	0.43	0.34	0.62
Fine-tuned on NewsQA-Entities	Greedy	**1**	**1**	0.38	0.40	**1**	0.50	0.44	0.66
	Beam search	3	0.57	0.62	0.63	0.54	**0.67**	0.64	**0.78**
	Top-k sampling	3	0.58	0.57	0.57	0.55	0.61	0.59	0.75
	Top-p sampling	3	0.56	0.58	0.58	0.53	0.62	0.59	0.75
	Marginal decoding	2.4	**0.93**	0.66	0.66	0.74	0.60	0.57	0.72
Fine-tuned on NewsQA-Entities + Uniform	Greedy	**1**	**1**	0.38	0.38	**1**	0.50	0.43	0.66
	Beam search	3	0.57	0.64	0.64	0.54	**0.67**	0.65	**0.78**
	Top-k sampling	3	0.57	0.59	0.59	0.56	0.61	0.58	0.75
	Top-p sampling	3	0.55	0.59	0.59	0.53	0.62	0.60	0.75
	Marginal decoding	2.8	**0.93**	**0.90**	**0.90**	0.71	**0.67**	**0.68**	**0.78**

While questions decoded with beam search also have a high overlap with the ground truth, they are less diverse than the ones produced by our approach. Closer examination of the results confirms that the QA pairs generated with beam search are paraphrases.

4.6 Results on Monserrate

In addition, we evaluate our approach on the manually produced benchmark dataset (see Table 3). Since Monserrate does not have answers annotated, we evaluated only diversity of the produced answers and measure both diversity and overlap with the ground truth for the generated questions.

The first thing to note is that our approach generates fewer QA pairs than were produced by human annotators. This is an expected result since not all of the questions in Monserrate are entity-based, however we still manage to match approximately a third of those questions. Marginal decoding again shows the best results in terms of producing diverse QA pairs that match the ground-truth data.

We manually inspected the generated QA pairs and concluded that all the entity-answers present in the dataset were identified correctly. The ground truth also contains the questions that address those entity-answers. Some of the generated questions, however, do not match the answers or the input sentence well. We dedicate the next section to a more systematic analysis that quantifies different types of errors our model makes.

5 Error Analysis

To further strengthen the results of our automated evaluation, we conduct manual annotation of the QA pairs produced by our approach. Since the automated metrics indicate the discrepancy between the produced sequences and the ground-truth sets, we use them to identify cases that deviate from the ground truth (with ROUGE score less than 0.5) and examine them in more detail.

The analysis was performed on a sample of the NewsQA-Entities dataset. We sample a set of answers that deviate from the ground truth using our softM0.5 metric and questions that have the same answers but differ according to softM0.5. Using a small subset of the results we first identify the most common error types and then employ a third-party team of professional annotators to perform an independent analysis of the produced results. We obtained 238 annotated answers and 201 annotated questions, in total. Results indicate that the majority of questions and answers are correct even when they deviate from the ground truth.

The majority of answers are either entities (44%) or entities with context spans surrounding those entities (13%). 43% of the answers are not entities but spans extracted from the input sentence. This result suggests that our model learned to identify new entities and potential answers that play a similar role in the sentence but cannot be detected by the original NER model. The most common mistake the model makes when generating the answers is by truncating them too soon (24%). 16% of the answers produced were not found in the input text and constitute proper mistake which is a known drawback of generative models.

The errors identified at the question generation phase are mostly the questions that do not have an answer in the input sentence (21%) or do not match the answer produced by the model (9%). These are the same error types, which the original model suffers from as well in similar proportions.

Table 4 provides examples of the generated QA samples alongside with the input sentences. It demonstrates that the baseline approaches are more prone to producing duplicates and question paraphrases than our approach.

6 Related Work

A recent survey on question generation provides a comprehensive overview of the research work alongside multiple dimensions [19]. It clearly points out that improving diversity of the generated questions as well as training with multi-reference data are bot important directions that are largely overlooked by the current approaches. This is the exact research gap which we are addressing with this work.

While multi-reference training for QAG was not previously studied in the related work, it has been already applied in other natural language generation tasks, such as machine translation and dialogue response generation. In [20], the authors generated pseudo-references for machine translation and image captioning task and then convert this multiple reference dataset into a single reference dataset for training the model. We consider this approach as our baseline (Fine-tuned on NewsQA-Entities).

To bootstrap a multi-reference dataset for training and evaluation of our model, we make use of a pre-trained NER model and many previously proposed QAG approaches utilised named entities. In [9] question templates are filled with entities from the input text when

Table 3: Evaluation results of our model fine-tuned on NewsQA-Entities with uniform training on the Monserrate corpus. #QA refers to the average number of QA pairs generated per input sentence.

Decoding	#QA pairs	Answers	Questions		
		Distinct-1	Distinct-1	ROUGE-L mean	softM0.5
Ground truth	9	-	0.47	**1**	**1**
Greedy	1	**1**	**1**	0.35	0.19
Beam search	4	0.51	0.44	0.42	0.30
Top-k sampling	4	0.52	0.46	0.42	0.30
Top-p sampling	4	0.49	0.45	0.41	0.27
Marginal decoding	4	**0.83**	**0.56**	**0.45**	**0.33**

Table 4: Examples of the generated QA pairs. Duplicate questions and answers are highlighted **in bold**.

Input sentence	Ground truth	Trained on NewsQA	Fine-tuned on NewsQA-entities with beam search	Fine-tuned on NewsQA-entities + uniform with marginal decoding
The service, named Fancast XFINITY TV (formerly called TV Everywhere), is the biggest cable industry initiative to keep people from skipping traditional TV service in the United States.	a> TV Everywhere) q> **What is the name of the new service?** a> Fancast XFINITY TV q> **What is the name of the new service?** a> the United States q> Where is the TV service?	a> **To keep people from skipping traditional TV service in the United States.** q> What did the program do? a> XFINITY TV q> What does the company want people to skip? a> **to keep people from skipping traditional TV service in the United States.** q> what is the fancast xfinity television?	a> Fancast XFINITY TV q> What is the name of the service? a> **the United States** q> Where is Fancast XFINITY TV? a> **the United States** q> Where is the service located?	a> Fancast XFINITY TV q> What is the name of the service? a> the United States q> Where is the TV service?
And also my time is important, so sitting down and watching a show at 8 or 9 p.m.	a> 8 or 9 p.m q> When is the show?	a> **8 or 9 p.m.** q> Time it's important what for him? a> And also my time is important, so sitting down and watching q> What did the girl says? a> **8 or 9 p.m.** q> When is that time a priority?	a> 8 or 9 p.m q> **When is the show?** a> 8 q> What time is important? a> 9 p.m q> **When is the show?**	a> 8 or 9 p.m q> When is the show?
Networks have tried for the past couple of years to find a way to reach Web-watching audiences by streaming content on their Web sites or making partnerships with Hulu, one of the larger online TV sites.	a> Hulu q> What is the name of the TV site? a> the past couple of years q> How long have networks tried to reach Web-watching audiences?	a> Hulu, one of the larger online TV sites. q> What companies are trying to do what with their content? a> to find a way to reach Web-watching audiences by streaming content on their Web sites or making partnerships with Hulu, q> What has the Nets tried? a> Networks q> Who has tried for years to reach Internet watchers?	a> **the past couple of years** q> **When did Networks try to reach Web-watching audiences?** a> **the past couple of years** q> How long have networks tried to reach Web-watching audiences? a> **the past couple of years** q> **When did the networks try to reach Web-watching audiences?**	a> Hulu q> What is one of the larger online TV sites? a> **the past couple of years** q> **When did Networks try to find a way to reach Web-watching audiences?** a> **The past couple of years** q> **When did Networks try to reach Web-watching audiences?** a> Networks q> Who has tried to reach Web-watching audiences? a> Internet q> What kind of audiences do networks try to reach? a> one q> What is Hulu?

designing an automatic tutoring systems. None of those approaches, however, considered training a QAG model end-to-end using entities as a multi-reference QA dataset.

7 Conclusion

We proposed a new approach for generating diverse question-answer pairs and validated it in an experimental evaluation. Our end-to-end QAG model dynamically identifies question-worthy context words pre-trained on named entities and uses them to condition subsequent question generation. Our results suggest that the proposed approach is successful in generating the same answer-entities as in the ground truth produced by a pre-trained NER model as well as in identifying new question-worthy phrases beyond named entities. A manual examination confirms that the generated questions are correct even when they differ from the ground-truth generated by the baseline.

We showed that it is possible to train a single model using data generated by a pipeline approach and replace this pipeline approach to reduce friction. In our case, the NER-based heuristic was used to demonstrate our proposed approach allowing us to avoid manual collection of a new multi-reference dataset. It is also straight-forward to extend our NER-based heuristic with other answer extractors or generators. However, our model can also be trained on human labeled data to produce QA pairs that cannot be generated by the pipeline approach. For example, a human-in-the-loop approach allows to use NER and other similar heuristics to bootstrap data collection but modify generated samples by selecting only useful questions.

Acknowledgments

We thank Gianni Barlacchi for sharing with us the pre-processed version of the NewsQA dataset with answers mapped to the original sentences from the news articles.

References

[1] Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou, 'Question generation for question answering', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 866–874. Association for Computational Linguistics, (2017).

[2] Angela Fan, Mike Lewis, and Yann N. Dauphin, 'Hierarchical neural story generation', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 889–898. Association for Computational Linguistics, (2018).

[3] Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu, 'Accelerating real-time question answering via question generation', *CoRR*, **abs/2009.05167**, (2020).

[4] Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang, 'CQG: A simple and effective controlled generation framework for multi-hop question generation', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 6896–6906. Association for Computational Linguistics, (2022).

[5] Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto, 'Chaincqg: Flow-aware conversational question generation', in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 2061–2070. Association for Computational Linguistics, (2021).

[6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, 'The curious case of neural text degeneration', in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, (2020).

[7] Robin Jia, Mike Lewis, and Luke Zettlemoyer, 'Question answering infused pre-training of general-purpose contextualized representations', in *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, eds., Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, pp. 711–728. Association for Computational Linguistics, (2022).

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov, 'Natural questions: a benchmark for question answering research', *Transactions of the Association of Computational Linguistics*, (2019).

[9] David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne, 'Generating natural language questions to support learning on-line', in *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*, pp. 105–114. The Association for Computer Linguistics, (2013).

[10] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho, 'Document expansion by query prediction', *CoRR*, **abs/1904.08375**, (2019).

[11] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning, 'Stanza: A Python natural language processing toolkit for many human languages', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (2020).

[12] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, 'SQuAD: 100,000+ questions for machine comprehension of text', in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (2016).

[13] Hugo Rodrigues, Eric Nyberg, and Luísa Coheur, 'Towards the benchmarking of question generation: introducing the monserrate corpus', *Lang. Resour. Evaluation*, **56**(2), 573–591, (2022).

[14] Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, Bill Byrne, and Adrià Gispert, 'Product answer generation from heterogeneous sources: A new benchmark and best practices', in *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pp. 99–110, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[15] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman, 'Newsqa: A machine comprehension dataset', in *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pp. 191–200. Association for Computational Linguistics, (2017).

[16] Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang, 'Adversarial domain adaptation for machine reading comprehension', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, eds., Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, pp. 2510–2520. Association for Computational Linguistics, (2019).

[17] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych, 'GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 2345–2360. Association for Computational Linguistics, (2022).

[18] Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Mo Yu, and Ying Xu, 'It is ai's turn to ask humans a question: Question-answer pair generation for children's story books', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 731–744. Association for Computational Linguistics, (2022).

[19] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng, 'A review on question generation from natural language text', *ACM Trans. Inf. Syst.*, **40**(1), 14:1–14:43, (2022).

[20] Renjie Zheng, Mingbo Ma, and Liang Huang, 'Multi-reference training with pseudo-references for neural translation and text generation', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3188–3197. Association for Computational Linguistics, (2018).

8. Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing

Bibliographic Information

Vera Provatorova, Samarth Bhargav, **Svitlana Vakulenko**, Evangelos Kanoulas: Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing. EMNLP (1) 2021: 10501-10510.

Robustness Evaluation of Entity Disambiguation Using Prior Probes: the Case of Entity Overshadowing

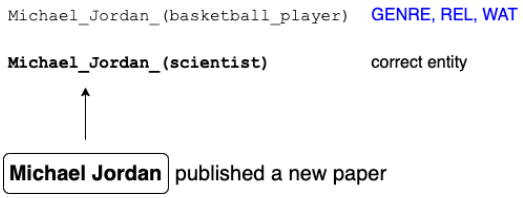
Vera Provatorova, Svitlana Vakulenko, Samarth Bhargav, Evangelos Kanoulas
University of Amsterdam, Amsterdam, The Netherlands
{v.provatorova, s.vakulenko, s.bhargav, e.kanoulas}@uva.nl

Abstract

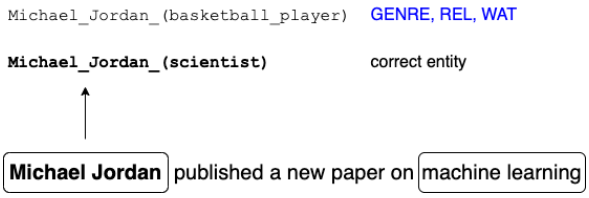
Entity disambiguation (ED) is the last step of entity linking (EL), when candidate entities are reranked according to the context they appear in. All datasets for training and evaluating models for EL consist of convenience samples, such as news articles and tweets, that propagate the prior probability bias of the entity distribution towards more frequently occurring entities. It was previously shown that performance of EL systems on such datasets is overestimated, since it is possible to obtain higher accuracy scores by merely learning the prior. To provide a more adequate evaluation benchmark, we introduce the ShadowLink dataset, which includes 16K short text snippets annotated with entity mentions. We evaluate and report the performance of several popular EL systems on the ShadowLink benchmark. The results show a considerable difference in accuracy between common and uncommon ambiguous entities that require disambiguation, for all of the EL systems under evaluation, demonstrating the effects of prior probability bias and entity overshadowing.

1 Introduction

The task of entity linking (EL) refers to finding named entity mentions in unstructured documents and matching them with the corresponding entries in a structured knowledge graph (Milne and Witten, 2008; Oliveira et al., 2021). This matching is usually done using the surface form of an entity, which is a text label assigned to an entity in the knowledge graph (van Hulst et al., 2020). Some mentions may have several possible matches: for example, "Michael Jordan" may refer either to a well-known scientist or the basketball player, since they share the same surface form. Such mentions are ambiguous and require an additional step of entity disambiguation (ED), which is conditioned on the context in which the mentions appear in the text, to be linked correctly. Following van Erp and Groth (2020) we refer to a set of entities that share the same surface form as an *entity space*.



(a) "Michael Jordan" (scientist) is overshadowed by "Michael Jordan" (basketball player).



(b) Even with more relevant context, overshadowing persists.

Figure 1: An example of entity overshadowing. The correct entity is ranked lower by the EL systems (indicated in blue) than the more common one.

To decide which of the possible matches is the correct one, an ED algorithm typically relies on: (1) contextual similarity, which is derived from the document in which the mention appears, indicating the *relatedness* of the candidate entity to the document content, and (2) entity importance, which is the prior probability of encountering the candidate entity irrespective of the document content, indicating its *commonness* (Milne and Witten, 2008; Ferragina and Scaiella, 2012; van Hulst et al., 2020).

The standard datasets currently used for training and evaluating ED models, such as AIDA-CoNLL (Hoffart et al., 2011) and WikiDisamb30 (Ferragina and Scaiella, 2012), are collected by randomly sampling from common data sources, such as news articles and tweets. Therefore, they are expected to mirror the probability distribution

with which the entities occur, thereby favouring more frequent entities (head entities) (Ilievski et al., 2018). From these considerations, we conjecture that the performance of existing EL algorithms on the ED task is overestimated. We set out to explore this effect in more detail by introducing a new dataset for ED evaluation, in which the entity distribution differs from the one typically used for training ED algorithms.

We perform a systematic study focusing on a particular phenomenon we refer to as *entity overshadowing*. Specifically, we define an entity e_1 as overshadowing an entity e_2 if two conditions are met: (1) e_1 and e_2 belong to the same entity space S, i.e., share the same surface form and, therefore, can be confused with each other outside of the local context; (2) e_1 is more common than e_2 in some corresponding background corpus (e.g. the Web), i.e., it has a higher prior probability $P(e_1) > P(e_2)$.

For example, e_1 = "Michael Jordan" (basketball player) overshadows e_2 = "Michael Jordan" (scientist) because $P(e_1) > P(e_2)$ in a typical dataset sampled from the Web. We use an unambiguous text sample that contains this mention to evaluate three popular state-of-the-art EL systems, GENRE (De Cao et al., 2020), REL (van Hulst et al., 2020), and WAT (Piccinno and Ferragina, 2014), and empirically verify that the overshadowing effect that we hypothesized, indeed, takes place (see Fig. 1a). Even when more information is added to the local context, including the directly related entities that were correctly recognised by the system ("machine learning"), the ED components still fail to recognise the overshadowed entity (see Fig. 1b).

The concept of overshadowed entities introduced in this paper is related to long-tail entities (Ilievski et al., 2018). However, these two concepts are distinct: a long-tail entity may be unambiguous and therefore not overshadowed, while an overshadowed entity may still be too popular to be considered a long-tail one.

To systematically evaluate the phenomenon of entity overshadowing that we have identified, we introduce a new dataset, called ShadowLink. [1]. ShadowLink contains groups of entities that belong to the same entity space. Following van Erp and Groth (2020), we use Wikipedia disambiguation pages to collect entity spaces. Disambiguation pages group entities that often share the same surface form and may be confused with each other. We then follow the links in the Wikipedia disambiguation pages to the individual (entity) Wikipedia pages to extract text snippets in which each of the ambiguous entities occur.

Note that we do not extract the text from these Wikipedia pages directly, since pre-trained language models such as BERT (typically used in state-of-the-art ED systems) also use Wikipedia as a training corpus, and can learn certain biases as well. Instead, we parse external web pages that are often linked at the end of a Wikipedia page as references. This data collection approach helps us to minimise the possible overlap between the test and training corpus.

Thereby, every entity in ShadowLink is annotated with a link to at least one web page in which the entity is mentioned. We then proceed to extract all text snippets in which the corresponding entity mention appears on the page. An extracted text snippet typically consists of the sentence in which the mention occurs.

Next, we use ShadowLink to answer the following research questions:

RQ1: How well can existing ED systems recognise overshadowed entities?

RQ2: How does performance on overshadowed entities compare to long-tail entities?

RQ3: Are ED predictions biased and how can we measure this bias?

Our contribution is twofold: (1) a new dataset for evaluating entity disambiguation performance of EL systems specifically focused on overshadowed entities, and (2) an evaluation of current state-of-the-art algorithms on this dataset, which empirically demonstrates that we correctly identified the type of samples that remain challenging and provide an important direction for future work.

2 The ShadowLink Dataset

This section describes the ShadowLink dataset: its construction process, structure, and statistics.

2.1 Dataset construction

The process of dataset construction consists of 3 steps: (1) collecting entities, (2) retrieving context examples for each entity, and (3) filtering the data based on the validity requirements detailed below.

Collecting entities. Similar to van Erp and

[1]ShadowLink dataset can be downloaded at https://huggingface.co/datasets/vera-pro/ShadowLink

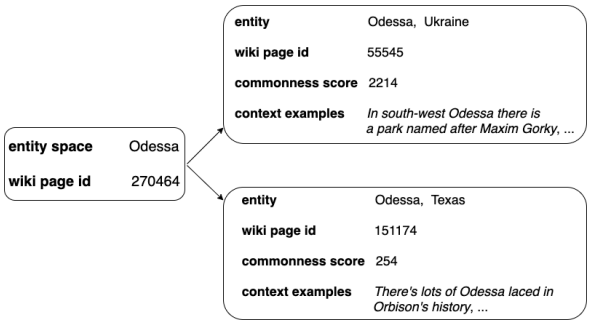


Figure 2: Structure of the ShadowLink dataset

Groth (2020), we use Wikipedia disambiguation pages to represent entity spaces. We retrieve a set of all Wikipedia disambiguation pages and filter it on the following criteria:

(1) For each disambiguation page (DP), we only include candidate entity pages with names containing the title of the DP as a substring. This step is required to exclude synonyms and redirects.

(2) If at least two candidate pages for the same DP match the criterion described above, then the DP and all its matching candidates are included as a new entity space.

During the first stage of the data collection, 170K out of 316.5K Wikipedia disambiguation pages matched the filtering criteria described above.

Filtering pages by year. To make sure that all pre-trained EL systems we evaluate in our experiments can potentially recognise all of the entities in the dataset, we also exclude pages that are more recent than the Wikipedia dumps used by these systems during training. The oldest dump used by a system in our experiments was the 2016 Wikipedia dump over which TagMe was trained, i.e we excluded all the pages that were created after 2016.

Collecting context examples. To retrieve context examples for each entity, we follow the external links extracted from the references section of the corresponding Wikipedia page and parse them to extract the text snippets which contain the entity mention. Then, every target entity mention is replaced with its corresponding entity space name, yielding an ambiguous entity mention. For example, if we have entities "John Smith" and "Paul Smith" that both belong to the entity space "Smith", then the mentions of both names will be replaced with "Smith". Looking for an entity name and replacing it with the corresponding entity space name (instead of looking for the entity space name in the first place) allowed us to make sure that the text snippets refer to the correct entity. Using this method, however, significantly reduced the number of retrieved snippets, as many of the entity mentions in natural texts do not include the full titles of the entities.

To extract the text snippets, we used a simple greedy algorithm that starts with the mention boundaries and tries to include more text, expanding the boundaries to the left and to the right, until it either covers one sentence on each side, or reaches the end (or beginning) of the document text. Our decision was to use relatively short spans similar to other popular ED benchmarks: WikiDisamb30 (Ferragina and Scaiella, 2012) and KORE50 (Hoffart et al., 2011). Our manual evaluation confirmed that these spans provide sufficient context for entity disambiguation. We also release the full-text of all web pages as part of our dataset, making the context of different lengths available for future experiments.

Commonness score. We estimate the commonness (popularity) of an entity as the number of links pointing to the entity page from other Wikipedia pages, that is, the in-degree of the entity page in the web graph of Wikipedia hyperlinks. Intuitively, this is proportional to the probability of encountering this entity when sampling a page at random. To obtain this metric for all the entities in the dataset, we use the Backlinks MediaWiki API[2].

Quality assurance. We conduct manual evaluation to assess the quality of the dataset and provide the upper bound performance for the ED task. The details of the setup and the results are discussed in Section 3.

2.2 Dataset structure and statistics

The ShadowLink dataset consists of 4 subsets: *Top*, *Shadow*, *Neutral* and *Tail*. The Top, Shadow and Neutral subsets are linked to each other through the shared entity spaces. On the other hand, the Tail subset, which contains (typically unambiguous) long-tail entities, is not connected to the other three through the same entity spaces. Nevertheless, it is collected in a similar way as the other three subsets.

Top and Shadow subsets. The structure of the Top and Shadow subsets is shown in Figure 2. Every entity e belongs to an entity space S_m, derived

[2] https://www.mediawiki.org/wiki/API:Backlinks

from the Wikipedia disambiguation pages, where m is an ambigous mention that may refer to any of the entities in S_m. Every S_m contains at least two entities: one e_{top} and one or more e_{shadow} entities. Every entity $e \in S_m$ is annotated with a link to the corresponding Wikipedia page and provided with context examples. A context example is a text snippet extracted from one of the external pages which contains the mention m , with a length of 25 words on average.

Neutral subset. To quantify the strength of the prior of each ED system, we synthetically generate data points for which the context around an entity mention is not useful for disambiguating that mention. To do that we use 7 hand-crafted templates. An example of such a template is the following: "It was the scarcity that fueled our creativity. This reminded me of m today." For each entity space, we generated 7 random contexts.

Tail subset. To evaluate the performance of ED systems on long-tail but typically not overshadowed entities, we collect an additional set of entities by randomly sampling Wikipedia pages that have a low commonness score ($<= 56$ backlinks)[3].

Context examples for these pages were collected in the same manner as described above. The resulting dataset matches the size and structure of other ShadowLink subsets, containing 904 entities.

The sampling process used to collect this subset follows the existing definition of long-tail entities(Ilievski et al., 2018), and is controlled for popularity but not for ambiguity. The Tail subset serves as a control group for the experiments conducted in our study, showing that the concept of entity overshadowing differs from the previously studied long-tail entity phenomena.

ShadowLink statistics. The dataset statistics across all the subsets are summarised in Table 1. Note that the *Top*, *Shadow* and *Neutral* subsets are grouped around the same entity spaces, while the *Tail* subset is constructed by sampling the same number of non-ambiguous entities. Every entity space contains at least 2 entities, with the mean number of entities per space being 2.63, median 2, and maximum 10. Figure 3 shows the distribution of commonness in the three subsets: Top, Shadow and Tail.

For the experiments we used a smaller subset of ShadowLink, with only one randomly selected shadow entity per entity space and one text snippet per entity. Thus, every subset contained 904 entities, with the total size of 9K text snippets. The rest of the data is left out as a training set and can be used in future experiments.

[3]This threshold is equal to the median number of backlinks in the *Shadow* subset.

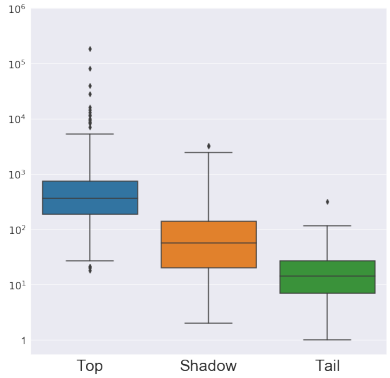


Figure 3: Distribution of the commonness score on the three subsets of ShadowLink.

3 Manual Evaluation

We perform manual evaluation of a random sample from ShadowLink to assess its quality, with the goal of ensuring that the extracted text snippets provide context sufficient for disambiguation. Human performance also sets the skyline for automated approaches on this dataset. In the following subsections, we describe the evaluation setup and the results of the manual evaluation.

3.1 Manual evaluation setup

We conduct a manual evaluation to assess the quality of the dataset and evaluate how well human annotators can disambiguate overshadowed entities. A sample of 91 randomly selected dataset entries was presented to two annotators, who examined the entries independently. For each entry, the annotators were presented with a text snippet containing an ambiguous entity mention m, and two entities, *Top* and *Shadow*, from the same entity space S_m, where one of the two entities was the correct answer. The annotators were instructed to either indicate the correct entity or mark the text snippet as ambiguous, which indicates that the provided context is not sufficient for the disambiguation decision to be made. Note, however, that the commonness scores were not displayed to the annotators.

Subset	# Entity Spaces	# Entities	# Text Snippets	Avg. # Words	Avg. # Sentences
Top	904	904	2K	29.25	1.11
Shadow	904	1.5K	6K	28.97	1.11
Neutral	904	-	6K	14.83	1.87
Tail	-	904	2K	28.94	1.10

Table 1: Dataset statistics across all the subsets of ShadowLink. The average number of words and sentences were calculated per text snippet extracted from the corresponding web page.

	Shadow	Top
	P = R = F	P = R = F
Annotator 1	0.973	0.973
Annotator 2	0.950	0.919
Average	0.963	0.946

Table 2: Results of the manual annotations.

3.2 Results of the manual evaluation

We used Cohen's kappa coefficient to evaluate the inter-annotator agreement (**?**) on all entries reviewed by the annotators. The value of the coefficient is 0.845, indicative of strong agreement. Next, we discarded the samples labelled as ambiguous by at least one of the annotators. The resulting dataset included 77 entries out of 91, which shows that 85% of the context examples were sufficient for making ED decisions. These unambiguous entries were split into two subsets, resulting in the 37 top-entities and 40 shadow-entities. We then discarded 3 randomly selected shadow-entities to achieve the same size of the two subsets, and used these subsets to evaluate the performance of manual ED for the top- and shadow-entities separately. The averaged F-score of the two annotators is 0.95 on the top-entities and 0.96 on the shadow-entities. The detailed results of the evaluation are shown in Table 2.

The results of manual evaluation show that (1) a majority of samples (85%) in ShadowLink are suitable for ED evaluation, i.e., automatically extracted snippets provide sufficient context for correct disambiguation; (2) human annotators can correctly disambiguate entities regardless of their commonness. Therefore, the performance of an automatic system that only depends on context is only bound by the 15% of the cases for which the context is not helpful. This bound can be further elevated if longer contexts are considered. Experiments on longer contexts are possible using the ShadowLink dataset[4] but we leave it for future work.

In the next section, we report and analyse the results produced by state-of-the-art systems on ShadowLink.

4 Benchmark Experiments

In this section, we describe the benchmark experiments designed to evaluate the baseline systems' performance on the ShadowLink dataset. For these experiments, we created a subset of the original dataset by sampling only one of the shadow entities at random to make the number of *Top* and *Shadow* equal. Note that in our task setup the model's predictions are not restricted to the top versus shadow entity binary decision. The model can predict any entity from the same or different entity space. We describe the experimental setup in Section 4.1, report the benchmarking results and analyse them in more detail in Section 4.2.

4.1 Evaluation setup

To answer the first two research questions (RQ1 & RQ2), we compare the performance of eight entity linking systems on the ShadowLink dataset. We used the GERBIL framework (Röder et al., 2018) for six of the baselines (AGDISTIS/MAG, AIDA, DBpedia Spotlight, FOX, TagMe 2 and WAT)[5] under the D2KB experimental setup[6]. We also

[4] We have crawled the full articles, and will be released as part of the ShadowLink datset.

[5] These six systems were the ones available on GERBIL at the time of our experiments.

[6] In the D2KB setup, the systems are provided with correct mention boundaries to evaluate the disambiguation step of entity linking.

performed an evaluation with the same setup using GENRE and REL, two novel state-of-the-art systems not available in GERBIL. We used micro-averaged precision, recall and F-score as evaluation metrics.

To answer the last research question (RQ3), we want to verify whether the baseline systems utilise context or simply rely on their priors to make the predictions. To this end, we compare the predictions made on the *Top*, *Shadow* and *Neutral* subsets. We used the predictions made on the *Neutral* subset as an indication of priors. That is, for each entity space, we generate context for the *Neutral* subset by using the same 7 random sentences as templates. The context was generated as neutral, i.e., it is not useful for the disambiguation task by design. Therefore, we considered the predictions for such neutral contexts to exhibit the default priors of an EL system for the given entity space. We can then compare these prediction to the predictions on the original examples from the *Top* and *Shadow* subsets. If the entity predicted for non-neutral context differs from the prediction made for the neutral context, we consider that the model updated its default prediction (prior) based on the local context. We performed this type of analysis to examine the predictions of the best-performing systems in our experiments: REL, GENRE, AIDA and WAT.

4.2 Benchmark Results

This section presents the results of our experiments and summarizes the answers to the research questions introduced in Section 1.

RQ1: *How well can existing ED systems recognise overshadowed entities?*

Table 3 shows the evaluation results across the subsets of ShadowLink. All systems achieve the lowest scores on the *Shadow* subset, with the maximum F-score of 0.35 achieved by AIDA. While REL and GENRE ourperform WAT on several existing datasets (van Hulst et al., 2020; De Cao et al., 2020), their results on ShadowLink are considerably lower than the results of WAT. The difference in the results on *Top* and *Shadow* entities indicates that EL predictions are biased towards more common entities.

RQ2: *How does the performance on overshadowed entities compare to long-tail entities?*

All systems show the highest precision on the *Tail* subset, i.e., they achieve much higher performance on the less ambiguous long-tail entities, compared to both top and overshadowed entities in ShadowLink. These results indicate that the main challenge in ED is the combination of ambiguity and uncommonness, while uncommon but non-ambiguous entities are relatively easy to resolve.

These findings are also consistent with Ilievski et al. (2018), who suggest that rare and ambiguous entities constitute the hardest cases for the EL task. In this study, we showed that such overshadowed entities indeed consititute a major challenge for the state-of-the-art systems and that ShadowLink provides a suitable benchmark for their evaluation.

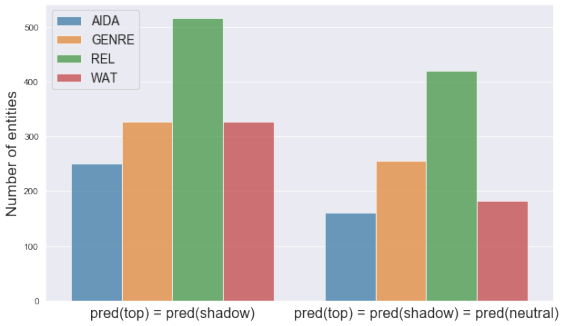


Figure 4: The degree of overshadowing (left) and prior bias (right) for each of the EL systems.

RQ3: *Are ED predictions biased and how can we measure this?*

Our experiments show that all systems under evaluation are often insensitive to the context change, i.e., the systems are actually unable to exploit local context for entity disambiguation but solely rely on their priors learned from the data. The error analysis results presented in Table 4 indicate that the majority of correct answers on the *Top* dataset coincide with the predictions observed on the *Neutral* subset. On the *Shadow* subset, opposite is the case: most of the errors are due to priors, and most of the correct predictions differ from them.

Figure 4 shows the number of cases in which overshadowing occurs for each of the systems, i.e., when the model's prediction remains the same for both *Top* and *Shadow* mentions. We see that this effect correlates with the number of cases in which the prediction of the system does not change regardless of the context, i.e., also for the *Neutral* context the prediction of the system remains the same. This observation confirms our initial hypothesis about the phenomena: the more common entities not only

Baseline	Shadow			Top			Tail		
	P	R	F	P	R	F	P	R	F
AGDISTIS/MAG (Usbeck et al., 2014)	0.14	0.14	0.14	0.25	0.25	0.25	0.79	0.79	0.79
AIDA (Yosef et al., 2011)	**0.40**	**0.31**	**0.35**	0.62	0.50	0.56	0.92	0.53	0.67
DBpedia Spotlight (Mendes et al., 2011)	0.16	0.10	0.12	0.41	0.25	0.31	**0.97**	0.11	0.19
FOX (Speck and Ngomo, 2014)	0.17	0.07	0.10	0.29	0.14	0.19	0.82	0.35	0.49
TagMe 2 (Ferragina and Scaiella, 2010)	0.34	0.25	0.29	0.69	0.49	**0.57**	0.95	0.74	0.83
WAT (Piccinno and Ferragina, 2014)	**0.40**	0.19	0.26	**0.72**	0.39	0.51	0.95	0.49	0.65
GENRE (De Cao et al., 2020)	0.26	0.26	0.26	0.42	0.42	0.42	0.93	**0.93**	**0.93**
REL (van Hulst et al., 2020)	0.21	0.21	0.21	0.54	**0.54**	0.54	0.91	0.91	0.91

Table 3: Benchmark evaluation results on the ShadowLink subsets.

	Top					Shadow				
	pred = prior		pred ≠ prior		NIL	pred = prior		pred ≠ prior		NIL
	correct	wrong	correct	wrong		correct	wrong	correct	wrong	
AIDA	15.5	12.9	35.0	28.0	8.6	4.8	20.0	28.2	39.3	7.9
WAT	32.1	11.7	16.4	12.4	27.4	4.2	31.8	20.9	15.6	27.5
GENRE	15.6	32.4	26.7	26.3	0.0	2.2	43.6	23.9	30.3	0.0
REL	32.3	24.8	21.4	21.0	0.6	7.7	50.9	12.9	27.9	0.6

Table 4: Error analysis, which shows the percentage of errors and correct predictions that either coincide (pred=prior) or differ (pred≠prior) from the predictions made for the neutral contexts, which we consider as predictions with the highest prior probability.

overshadow the less common ones but they are also used as the default predictions made completely independent of the given context, which we call the system priors.

Figure 4 shows that among the four best EL systems, REL is the most prone to overshadowing and prior bias. This also explains its poor performance on the $Shadow$ subset in comparison with the high performance demonstrated on $Tail$. AIDA and WAT appear to be more sensitive to the local context, which allows them to achieve better results on the overshadowed entities in comparison to both GENRE and REL. Moreover, AIDA, which outperforms all other systems on the *Shadow* subset, turns out to be the least affected by the overshadowing phenomena. These results indicate that the main reason behind the poor ED performance on overshadowed entities is due to systems overrelying on the prior bias and failing to incorporate contextual information.

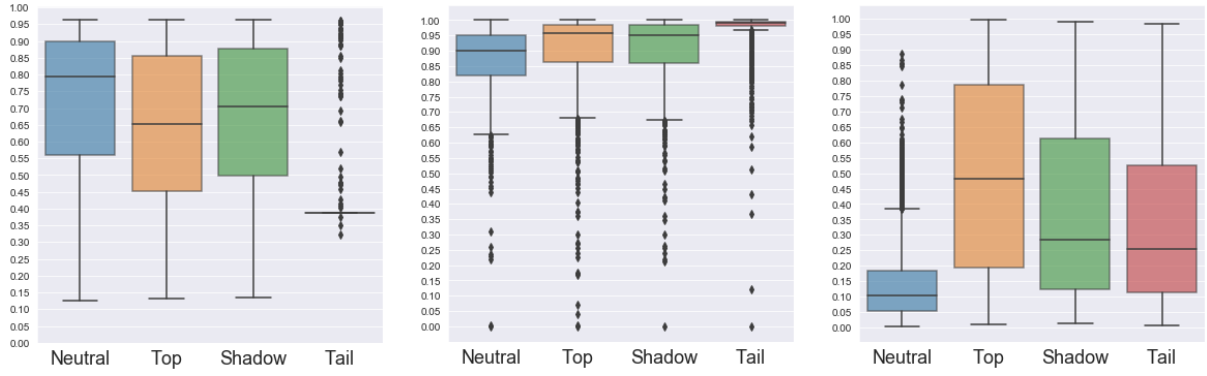
Lastly, we also look at the confidence scores for each of the subsets to check if they can be used as an additional indicator (see Figure 5). Interestingly, the systems have very different distributions of their confidence scores. For example, WAT has lower confidence when given neutral samples, which can be used to detect context ambiguity and filter out such samples. However, this approach can not be used for REL's and GENRE's predictions[7].

5 Related Work

Datasets for ED evaluation. Evaluation of ED performance was on the research radar for several years, and many benchmark datasets were proposed to date (Hachey et al., 2013; Röder et al., 2018; Ehrmann et al., 2020). Among the most popular ones are AIDA-CONLL (Hoffart et al., 2011), which consists of 1.4K annotated news articles with 27.8 entity mentions; AQUAINT dataset (Milne and Witten, 2008) with 50 news articles and 727 mentions; MSNBC (Cucerzan, 2007) with 20 news articles and 656 mentions. However, the standard benchmarks used for ED evaluation do not reflect the challenges that are often encountered in practice, such as limited context, long-tail, emerging and complex entities (Meng et al., 2021).

Guo and Barbosa (2018) construct two datasets by sampling hard ED examples from Wikipedia and

[7]GENRE's confidence scores were rescaled before the comparison.



(a) REL (b) GENRE (c) WAT

Figure 5: Distribution of confidence scores on all subsets of ShadowLink.

Aspect	ShadowLink	WikiDisamb30	KORE50
Source	Web	Wikipedia	Manual
Long-tail entities	✓	-	✓
# Mentions	15K	1.4M	148

Table 5: ShadowLink in comparison with other datasets that specifically focus on the entity disambiguation task.

ClueWeb corpora on which a simple baseline using priors does not succeed. Their experiments show that this prior-based baseline achieves a high performance, which also indicates the need for more challenging evaluation datasets. ShadowLink aims to close this gap. In this work, we focus specifically on the long-tail entities since the existing benchmarks are known to be biased towards the head of the distribution, i.e., the popular entities (Ilievski et al., 2018; Guo and Barbosa, 2018).

Similarly to ShadowLink, WikiDisamb30 (Ferragina and Scaiella, 2012) contains short text snippets annotated with Wikipedia entities designed for ED evaluation. In contrast to WikiDisamb30, the text snippets in ShadowLink were extracted from web pages outside of Wikipedia to avoid the effects of overfitting since Wikipedia is often used for training language models. Moreover, ShadowLink examples were collected using Wikipedia disambiguation pages as entity spaces while WikiDisamb30 represents a random sample from Wikipedia that does not allow to examine the effect of overshadowing.

The idea of entity spaces was previously introduced by van Erp and Groth (2020), who showed that predicting entity spaces largely improves recall. Their results also hint on the conclusion that disambiguation within entity spaces constitutes a bottleneck in the ED performance. We take this idea further by designing a dataset centered around entity spaces to evaluate ED performance within entity spaces directly. This dataset allows us to measure the gap the state-of-the-art ED systems still have on this task.

KORE50 (Hoffart et al., 2012) was created to evaluate the impact of low commonness and high ambiguity on the ED performance but it contains only 50 hand-crafted sentences with 148 entity mentions including ambiguous mentions and long-tail entities. ShadowLink continues this line of work, providing a considerably larger number of samples that can be used for training and evaluation of ED approaches. We also introduce a subset of neutral samples designed to uncover the model priors. Table 5 summarizes how ShadowLink differs from the previously introduced datasets for entity disambiguation.

Robustness evaluation. Our approach to ED evaluation taps into the fast-growing area of research aimed at assessing model robustness especially relevant for data-driven machine learning techniques. One of the first studies on this topic (Sturm, 2014) argued that the state-of-the art music information retrieval systems show very good performance on the standard benchmarks without the real understanding of the task at hand since their predictions relied solely on the confounds present in the ground truth. Sturm (2014)

also coined the term for this phenomena: the "Clever Hans" effect, named after the infamous horse that appeared to solve arithmetic problems while only following unintentional body language cues given by the trainer. More recently, **?** showed that the same effect is demonstrated by other state-of-the-art machine learning models, and the standard performance evaluation metrics fail to detect it. **?** further explored this phenomenon, showing that it also affects the reliability of unsupervised models in the field of anomaly detection. Therefore, not surprisingly we also observed this effect in the ED task: Guo and Barbosa (2018) used a rudimentary system that merely learned the prior distribution of entities to disambiguate them, and demonstrated that it performs on par with state-of-the-art approaches. These findings specifically calls for new datasets that allow for a more robust evaluation and deeper analysis of the model performance, similar to the one demonstrated here with ShadowLink. We hope that this paper might inspire similar datasets in other fields, where the priors from large public datasets may also overshadow the local context.

6 Conclusion

We introduced ShadowLink, a new benchmark dataset for evaluating entity disambiguation performance, and used it for an extensive analysis of the state-of-the-art systems' results. Our experimental results indicate that all systems under evaluation are prone to rely on their priors, which explains their higher performance on more common entities, and much lower performance on the lexically similar overshadowed entities. Our work thereby shows that the ED task is still far from solved for overshadowed entities, and ShadowLink paves the way for further research in this direction.

The shortcomings of existing disambiguation approaches uncovered by the ShadowLink dataset stimulate further research towards developing more robust ED algorithms that are better at exploiting context without overrelying on the prior bias. We would also like to explore ways to account for more context around the entity mentions, and when expanding the context is actually needed.

Acknowledgements

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the Informatics Institute of the University of Amsterdam, the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the Google Faculty Research Awards program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria. INCOMA Ltd.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 708–716.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, pages 288–310.

Marieke van Erp and Paul Groth. 2020. Towards entity spaces. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2129–2137, Marseille, France. European Language Resources Association.

Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1625–1628.

Paolo Ferragina and Ugo Scaiella. 2012. Fast and accurate annotation of short texts with wikipedia pages. *IEEE software*, 29(1):70–75.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R Curran. 2013. Evaluating entity linking with wikipedia. *Artificial intelligence*, 194:130–150.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. KORE: keyphrase overlap relatedness for entity disambiguation. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 545–554.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. REL: an entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2197–2200.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011*, pages 1–8.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gemnet: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021 (To appear)*.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Italo L. Oliveira, Renato Fileto, René Speck, Luís P.F. Garcia, Diego Moussallem, and Jens Lehmann. 2021. Towards holistic entity linking: Survey and directions. *Information Systems*, 95:101624.

Francesco Piccinno and Paolo Ferragina. 2014. From tagme to WAT: a new entity annotator. In *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*, pages 55–62.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625.

René Speck and Axel-Cyrille Ngonga Ngomo. 2014. Named entity recognition using fox. In *International Semantic Web Conference (Posters & Demos)*, pages 85–88. Citeseer.

Bob L Sturm. 2014. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. Agdistis - agnostic disambiguation of named entities using linked open data. In *ECAI*, volume 2014, pages 1113–1114. Citeseer.

Mohamed Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. 2011. Aida: An online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4:1450–1453.

9. VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law

Bibliographic Information

Julien Rossi, **Svitlana Vakulenko**, Evangelos Kanoulas: VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law. CIKM 2021: 4554-4563.

VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law

Julien Rossi
j.rossi@uva.nl
University of Amsterdam
Netherlands

Svitlana Vakulenko
s.vakulenko@uva.nl
University of Amsterdam
Netherlands

Evangelos Kanoulas
e.kanoulas@uva.nl
University of Amsterdam
Netherlands

ABSTRACT

Citing legal opinions is a key part of legal argumentation, an expert task that requires retrieval, extraction and summarization of information from court decisions. The identification of legally salient parts in an opinion for the purpose of citation may be seen as a domain-specific formulation of a highlight extraction or passage retrieval task. As similar tasks in other domains such as web search show significant attention and improvement, progress in the legal domain is hindered by the lack of resources for training and evaluation. This paper presents a new dataset that consists of the citation graph of court opinions, which cite previously published court opinions in support of their arguments. In particular, we focus on the verbatim quotes, i.e., where the text of the original opinion is directly reused. With this approach, we explain the relative importance of different text spans of a court opinion by showcasing their usage in citations, and measuring their contribution to the relations between opinions in the citation graph. We release VerbCL[1], a large-scale dataset derived from CourtListener and introduce the task of highlight extraction as a single-document summarization task based on the citation graph establishing the first baseline results for this task on the VerbCL dataset.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; *Content analysis and feature selection*; *Summarization*; **Expert search**; *Information extraction*; *Question answering*; • **Computing methodologies** → *Information extraction.*

KEYWORDS

case law, information retrieval, summarization

ACM Reference Format:
Julien Rossi, Svitlana Vakulenko, and Evangelos Kanoulas. 2021. VerbCL: A Dataset of Verbatim Quotes for Highlight Extraction in Case Law. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3459637.3482021

[1]https://github.com/j-rossi-nl/verbcl

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482021

1 INTRODUCTION

Courts of common law systems rely on statutes and precedents for deciding the law applicable to a case. A court opinion on a case, being a decision taken by a court when litigating a specific case, is a text assembled by the court judges using the parties' arguments and rebuttals of arguments. Previously published opinions (also known as case law) are cited to support the argumentation of the parties and the opinion of the court on the litigated case.

Thereby a legal discourse is made of a mix of factual details from the case as well as legal discussions arising from the qualification of the facts, driven by the past discussions with a similarity, either in facts or in reasoning. We provide an example, taken from the VerbCL dataset that illustrates how case law citation is used in practice, in Table 1.

For any legal-domain practitioner, search for an appropriate citation is a key task in preparing the argumentation for a case, which needs to be facilitated by an information retrieval system. This task was considered in case law retrieval [18, 33] as an information retrieval task where the query is a draft of an opinion and the result is a ranked list of past opinions. Relevance in this case may be guided by the similarity between the query and the retrieved opinion, or by the fact that the retrieved opinions address legal facets or questions in the query, as well as by the consideration whether the retrieved opinions might be worth being cited or not in the context of the query.

In this work, we introduce the new task of *highlight extraction*: given the text of an opinion, predict the subset of text spans that are likely to be cited in the future. In this task, the question is whether we can successfully predict which parts of an opinion will have an impact on future litigation, i.e., will be cited.

The motivation behind this task is to facilitate the work done by a legal commentator, who is tasked to identify interesting points in an opinion at the time of its publication. The legal commentator inspects all recently published opinions looking for potentially valuable excerpts that are worth being cited. The evaluation whether a part of an opinion is of a particular interest, or might be of interest in future cases is informed by the knowledge of the jurisprudential landscape, so the commentator can contrast in a new opinion what makes part of this landscape, and what is novel.

We aim to inform the task of highlight extraction through the construction of a large-scale dataset that will allow the development and evaluation of data-driven models able to solve the task. To this end, we use a public dataset of US court opinions, CourtListener[2], to extract a citation network of court opinions.

The citation network naturally emerges from the repository of opinion documents since each opinion usually quotes multiple

[2]https://www.courtlistener.com

Citing opinion	**ID**: 2346822 **Title**: LL Ex Rel. Doe v. Chimes Dist. of Columbia, Inc **Text**: (...) a special relationship that gave rise to a duty on the part of the United States to protect or warn her.[1] There was no such special relationship between L.L. and the United States, however, and the United States motion must accordingly be granted. "Ordinarily, the owner or possessor of land is under no duty to protect invitees from assaults by third parties while the invitee is upon the premises ... [unless] there is a special relationship between [the] possessor of land and his invitee giving rise to a duty to protect the invitee from such assaults." Wright v. Webb, CITATION_1239944 , 920-21 (1987). Chimes submits that this is a "special relationship" case. The cases on which Chimes relies for that proposition presented distinct factual bases for finding a "special relationship." (...)
Cited opinion (score = 1.71)	**ID**: 1239944 **Title**: Wright v. Webb **Text**: (...) We will assume, without deciding, that Webb was the Wrights' business invitee. Thus, the Wrights owed Webb the duty of ordinary care to maintain their parking lot in a reasonably safe condition. See Tate Rice, 227 Va. 341, 345, 315 S.E.2d 385, 388 (1984). Ordinarily, the owner or possessor of land is under no duty to protect invitees from assaults by third parties while the invitee is upon the premises. Restatement (Second) of Torts | 314A (1965) recognizes exceptions to the rule of non-liability for the assaults of a third party where there is a special relationship between a possessor of land and his invitee giving rise to a duty to protect the invitee from such assaults. We alluded to this Restatement rule in both Klingbeil Management Group Co. Vito, 233 Va. 445, 447, 357 S.E.2d 200, 201 (1987) and Gulf Reston, Inc. Rogers, 215 Va. 155, 158, 207 S.E.2d 841, 844 (1974), but made it plain in Gulf Reston that this was only a reference to the Restatement rule. Our statement in Klingbeil was simply a comment upon the reference in Gulf Reston. (...)

Table 1: A sample of a verbatim quote from the VerbCL dataset.

previously published opinions for the purpose of the legal argumentation of the case. We consider this network as a directed acyclic graph, where each node stands for an opinion document. Edges represent the citations made in opinions, directed from the citing opinion's node towards the cited opinion's node. In this paper, we consider the citing-cited relation as follows: a citing opinion is making an argument by referring to a cited opinion.

In the text of the citing opinion, a citation is introduced by an *anchor*, that we consider to be the supported legal argument. More specifically, we differentiate between two types of anchors: abstractive anchors, where the argument is cited as a paraphrase of the original opinion document, and extractive anchors (or *verbatim quotes*) where the argument is directly extracted from the cited opinion's text by copying the relevant text span. We focus on the verbatim quotes in this work since they are much easier to detect and reproduce than paraphrases.

We further consider that verbatim quotes emphasize which part of the cited opinion have a legal weight, bringing novelty and importance into the legal landscape at the time of the publication of an opinion document. We define a *highlight* of an opinion as the set of all text spans that are later cited as verbatim quotes. Therefore, we can produce the highlights for the previously published opinions by considering the citation graph. The highlight of an opinion can be constructed by considering all the verbatim quotes of this opinion used in the citing opinion documents (see an example of the highlight in Table 1).

To aggregate information about all citing opinions, the citation graph is used. We illustrate the citation graph around the cited opinion from Table 1 in Figure 1, which shows all the citing opinions as well as the identifiers of their verbatim quotes as the edge labels.

The goal of highlight extraction is to predict the highlights for new opinions, i.e., predicting the future citations based on the opinion text. We proceed by establishing the first baseline for the highlight extraction task using the state-of-the-art single-document summarization approaches: TextRank [1] and PreSumm [27], as well as a sentence-level classifier based on DistilBERT [36] . The goal of our experiments is to verify to which extent the highlight extraction task in the citation network can be addressed using single-document summarization approaches without any information about the citation graph structure.

Our main contributions are two-fold:

(1) the dataset with a citation graph, verbatim quotes and opinion highlights;
(2) the task of highlight extraction with baselines.

We start by describing how VerbCL was constructed in Section 2 and discuss its main characteristics in Section 3. We formalize the task of highlight extraction and report on the experiments we conducted with the baseline models for this task in Section 4. The overview of existing datasets for case law retrieval, and related work on highlight extraction and citation-based summarization in other domains is given in Section 5. We conclude and propose directions for future work in Section 6.

2 DATASET CONSTRUCTION

VerbCL is based on the Court Listener dataset, a collection, which is publicly available for free. In this section, we describe the process of constructing the dataset.

Field Name	Type	Comment
`citing_opinion_id`	`int`	The `opinion_id` of the **CITING** opinion.
`cited_opinion_id`	`int`	The `opinion_id` of the **CITED** opinion.
`sentence_id`	`int`	The `sentence_id` of the source sentence in the CITED opinion.
`verbatim`	`str`	The span of text in the CITING opinion that we consider to be a potential verbatim quote from the CITED opinion
`snippet`	`str`	The span of text from the CITING opinion around the in-text citation of the CITED opinion. It contains 100 words before the in-text citation, and 100 words after the in-text citation.
`score`	`float`	The score of `verbatim` for being an actual verbatim quote from the CITED opinion, placed in the CITING opinion. A score of -1 indicates our model does not consider it as being a quote from the CITED opinion.

Table 2: Fields of the documents in VerbCL Citation Graph.

Field Name	Type	Comment
`opinion_id`	`int`	The unique identifier of the opinion.
`sentence_id`	`int`	The unique identifier of the sentence: its index within the list of sentences of the opinion.
`raw_text`	`str`	The complete text of the sentence.
`highlight`	`bool`	The binary label of the sentence. *True* means the sentence has been quoted verbatim in a citing opinion.
`count_citations`	`int`	The number of times the sentence has been quoted verbatim.

Table 3: Fields of the documents in VerbCL Highlights.

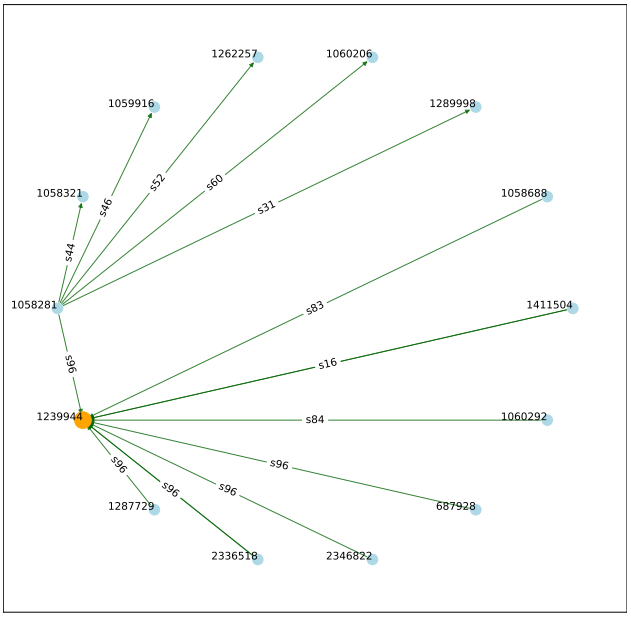


Figure 1: Citation graph, where each node is an opinion document with edges pointing from a citing opinion to a cited opinion. Every edge is labeled with the identifier of a sentence of the cited opinion that was quoted verbatim. For example, sentence 96 of opinion 1239944 is cited by the opinion 1058281, which contains sentence 44 cited by opinion 1058321.

2.1 Court Listener

CourtListener is a project of the Free Law Project[3], a US non-profit organization started in 2010, with the aim of "making the legal world more fair and efficient".

The original Court Listener dataset is a collection of every court opinion published by every court in the United States. It covers 406 jurisdictions (out of 423), with opinions from the year 1754 up to now. It is constantly updated with newly filed opinions, and digitized archives.

We obtained the Court Listener dataset by downloading the bulk data on September 1st 2019.[4]. More recent versions will have more opinions available, the whole processing will apply as long as the underlying relational database scheme remains unchanged.

The unit of data in Court Listener is the court opinion. Each court opinion is a single unique decision made by a specific court at a specific date on a specific case. A single case can be litigated by many courts, in case of appeals for example, so opinions are clustered per case. The Court Listener dataset follows the scheme of a relational database, where each opinion belongs to one cluster. Opinions are also clustered in dockets, a single docket gathering the different opinions made by a unique court on a unique case, as it happens that the same court may emit multiple opinions on the case at different dates.

At a more granular level, Court Listener represents each opinion as an HTML document, where specific tags mark where an in-text citation of case law appears in the opinion. Free Law implements a citation parsing and matching program[5] that allows for an automated annotation of the unstructured text of the original court opinion.

[3]https://free.law/
[4]https://www.courtlistener.com/api/bulk-info/
[5]https://github.com/freelawproject/courtlistener

2.2 Data Processing

For the construction of VerbCL, we focus on the opinions of CourtListener whose full text is available as HTML code including tagged in-text citations (looking for the field `html_with_citations`). The HTML code for each opinion has specific tags to identify citations made in the text of the opinion, each citation being identified with the unique identifier of the cited opinion.

We consider that each citation is introduced by an anchor, which is located in the vicinity of the citation tag. The anchor is the "raison d'être" of the citation, it is an argument to the current case, it is participating to the litigation of the case, and its validity is affirmed by invoking a similar argument that has already been made in an existing court opinion.

We focus specifically on the anchors that are extracts of the text of the cited opinion (*verbatim quotes*), similar to the example provided in Table 1.

Algorithm 1: Data Processing Pipeline

```
Result: Verbatim Quotes
initialization;
forall opinions do
    forall citations do
        snippet = N words before / after in-text citation;
        candidates = potential verbatim quotes;
        forall candidates do
            if candidate is verbatim then
                identify original sentences in cited opinion;
                store data point;
```

Our data processing procedure includes extracting the text around the in-text citation, identifying potential verbatim quotes (often marked by quotation marks) and eventually classifying those who actually are verbatim quotes from the text of the cited opinion. Our code includes a framework for parallel distributed processing of massive dataset, using `pyarrow`, Elasticsearch[6] through its Python API, and MongoDB[7] together with `pymongo`.

2.2.1 Snippets extraction. Although the anchor is in the vicinity of the in-text citation, we observe a wide range of variations in the presentation and wording of these anchors, in relation to the in-text citation. Our strategy is to reduce this problem to identify a verbatim quote in the direct surroundings of the in-text citation, defined as the N words before and after the citation. We opted for a value $N = 100$, which is a compromise between high recall (the more text around the citation we consider, the more likely it is we will identify the anchor), and computational footprint. Experiments included: considering entire paragraphs, but it yielded too much text; identifying nearby sentences, but the high numbers of acronyms and abbreviations around in-text citations renders sentence splitting useless.

2.2.2 Verbatim candidates. The identification of potential verbatim quotes is entirely rule-based, around the usage of different quote characters. Each snippet will generate multiple spans that were enclosed between those quote characters.

2.2.3 Qualifying verbatim quotes. At this stage, we want to confirm for each verbatim candidate whether it is a text that originates from the cited opinion. In the context of studying a specific in-text citation in a citing opinion, the cited opinion is known, so we have to evaluate whether a text span is lifted from the full text of the cited opinion.

The problem of identifying segments of text within a long text has proven to be difficult. As most of the opinions originate from printed books, OCR artifacts are expected, such as misplaced spaces or misspellings. This will affect the quality of the text in both cited and citing opinion. The common practice of using ellipsis in the citing opinion's text (as can be seen in the sample shown in Table 1) will result in the suppression of fragments of the original text.

Traditional approaches to this fuzzy matching problem rely on heuristics based on the computation of editing distances, such as the Levenshtein Distance [21]. The design of the heuristics is a first challenge, given the sheer variety of differences that could be observed between the original text in the cited opinion and its counterpart in the citing opinion. More importantly, the computational footprint of calculating an editing distance renders this method inapplicable to our dataset.

We built a classifier based on "Interval Queries" offered by Elasticsearch. We refer the reader to the provided source code for the exact parameters of this query. The classifier predicts that a span of text belongs to the positive class when the query returns a non-empty result, and predicts the negative class when this query does not produce any search result. We manually annotated a test dataset for this classifier which revealed it had high precision and recall for both classes. The reader can refer to Chapter 2.3.

As a result of this stage, we have built the VerbCL citation graph, a directed graph where nodes are opinions and edges are verbatim quotes.

2.2.4 Identifying highlights. Highlights in the cited opinion are the sentences that were used for verbatim quotes. For each span of text identified as a verbatim quote of the cited opinion, we identify which sentence or sentences, are actually quoted from the text of the cited opinion. This is achieved through another set of Elasticsearch queries that are run over an indexed collection of opinion sentences. For sentence tokenization, we used the Punkt sentence tokenizer from `NLTK` [2].

As a result, we have built the VerbCL Highlights dataset, a dataset of opinions annotated at sentence level with a binary label where the positive class is made of sentences that were later on cited as verbatim quotes.

2.2.5 Dataset structure. Following this stage, we have produced the following data collections:

- **VerbCL Highlights**: An annotated collection of all opinion sentences, with a binary label for which True indicates that the sentence was cited verbatim;
- **VerbCL Citation Graph**: An annotated citation graph for all case law citations with a verbatim quote of the cited opinion. It is a directed acyclic graph (V, E) where V is the

[6] https://www.elastic.co/
[7] www.mongodb.com

set of all opinion documents and E is the set of verbatim quotes.

VerbCL Highlights is used to inform the highlight extraction task since it contains the original opinion text and the subset we consider as the correct highlight. VerbCL Citation Graph was used to produce VerbCL Highlights and provides auxiliary information about the opinion text reuse that can be used for the highlight extraction task.

See Tables 2-3 for the structure of these subsets. A tutorial on loading the data is available as a `Jupyter Notebook` in the code repository[8].

2.3 Quality Assurance

We manually annotated 180 random anchors, coming from random opinions, and evaluated whether or not the candidate snippet was an actual verbatim quote out of the cited opinion. The code for this manual annotation task is included in our codebase, it is based on the `Django` framework.[9]

On this random sample of 180 texts, with 60 of them from the positive class (actual verbatim quotes from the cited opinion), our binary classifier had a Precision $P > 0.96$ and Recall $R > 0.92$ for both classes. We consider our solution for this problem to offer a sufficiently good compromise between computation and accuracy.

3 DATASET STATISTICS

In the following we describe the main characteristics of the VerbCL Citation Graph and the VerbCL Highlights.

3.1 VerbCL Citation Graph

From Court Listener, we consider the subset of opinions that either cite other opinions with a verbatim quote, or are cited verbatim by other opinions. This subset contains circa 1.5M opinions.

The main characteristics of our dataset are summarized in Table 4. Considering that we are studying the nodes of a citation network, we also disclose our analysis of the corresponding graph.

In the citation network, nodes are opinions and directed edges materialize the citations, from the citing opinion towards the cited opinion. Thereby, the in-degree of a node v is the number of times this opinion is cited by other opinions, while the out-degree indicates the number of opinions that the opinion corresponding to the node v cites.

We plot the distribution of in-degree of the nodes in the citation network in Figure 2, as well as the Zipf's law, a power law with an exponent $k = -1$, which is a common distribution in the field of text mining. This shows us that these verbatim quotes have a behavior closer to the appearance of a term in a collection of documents, than to the expected distribution of links in a web graph (a power law with exponent $k \approx 2.1$).

We imported the citation network, using `networkx` [12], in order to compute some descriptive centrality statistics of the graph. For computing time reasons, we used approximations instead of the actual values, following the experiments in [4].

[8] https://github.com/j-rossi-nl/verbcl
[9] https://www.djangoproject.com/

Court Listener (CL)

Opinions	4,265,231
Citing opinions	3,062,334
Cited opinions	2,020,779
Citations	30,318,321

VerbCL Citation Graph

Opinions	1,493,561
	35% of opinions in CL
Verbatim quotes	6,210,703
	20% of citations in CL
Verbatim citing opinions	1,086,238
	35% of citing opinions in CL
Verbatim cited opinions	946,962
	47% of cited opinions in CL
Edges in the citation graph	4,002,137
Graph density	1.8×10^{-6}
Number of words in a verbatim	$[5 - 100]$
Average	15
Quartiles 25-50-75	12 - 20 - 30

Table 4: Dataset statistics of Court Listener (CL) and VerbCL.

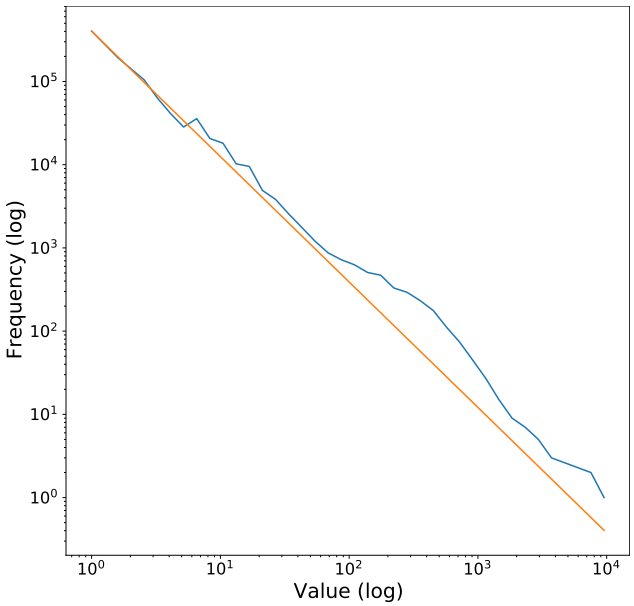


Figure 2: Distribution of number of citations for an opinion.

The distribution of betweenness centrality of the nodes of the citation graph is plotted in Figure 3, where we observe that the network contains a few very central nodes, corresponding to the group with the highest values. A node with a high betweenness centrality is a node which is on the shortest path between many nodes of the graph, reflecting its importance in the network. Betweenness centrality [3] of a node v is defined as the sum of the fraction of all-pairs shortest paths that pass through v, see Equation 1. Since the exact calculation of betweenness centrality for every node of

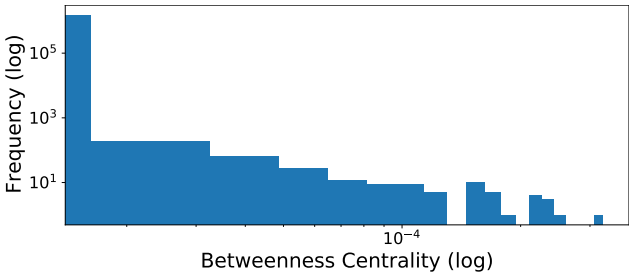


Figure 3: Distribution of betweenness centrality.

the network is too computationally expansive, we made use of an approximation algorithm [4].

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (1)$$

When applied to academic literature and the corresponding citation network, betweenness centrality is considered to measure importance of a paper in the field, i.e., seminal academic papers were reported to have a relatively high betweenness centrality [22, 31]. In contrast, our citation network has only a few nodes with high relative betweenness centrality but the absolute values are still very low (maximum: 3×10^{-4}).

We hypothesize that this difference can be explained by much higher content redundancy in the court opinions in comparison with academic literature. Novelty and originality of contributions are the main characteristics for an academic paper, which makes the paper take a unique and irreplaceable place in the network. In contrast, courts of the justice system serve any valid case presented regardless of its originality. Therefore, it is only natural that multiple cases can often make for equally good candidates in support of the same argumentation line. This has an effect of the more even citation distribution across the whole network of court opinions in comparison with an academic network, which also reduces the centrality measures for court opinions as our data confirms.

Comparing the ranking of nodes by centrality with the ranking by outdegree (i.e. the number of times this opinion is cited by others), we compute the Kendall's Tau and Spearman's Rho coefficients of rank correlation: $\tau = 0.21$, $\rho = 0.27$, both with p-value $p = 0.0$, showing weak but significant correlation. We attribute this weak correlation to the historicity factor: contrary to a webpages network, each node in our citation network is constrained by the capacity to cite only past opinions; we also consider the novelty factory, where a legal practitioner is biased towards most recent opinions when citing, therefore pushing the "central" opinions out of the shortest paths.

3.2 VerbCL Highlights

The main characteristics of our dataset are summarized in Table 5. These statistics come from a random sample of circa 10k opinions.

Sentences from one opinion that are cited verbatim by other opinions are considered as highlights of the cited opinion. The task of identifying the highlights from the full text of an opinion is introduced in this paper under the name "Highlight Extraction", in Section 4.

VerbCL Highlights

Opinions	1,493,561
Number of sentences per opinion	99.5% under 5000 sentences
Average	437
Number of tokens per sentence	99.9% under 171 tokens
Average	24
% highlight sentences (per opinion, the number of highlights divided by the number of sentences)	99.5% under 21.8%
Average	3.5%
Number of citations per highlight	99% under 434 citations
Average	34

Table 5: Dataset statistics for VerbCL Highlights.

Court opinions are typically large documents, we can illustrate this by showing the distribution of the number of words per opinion in Figure 4.

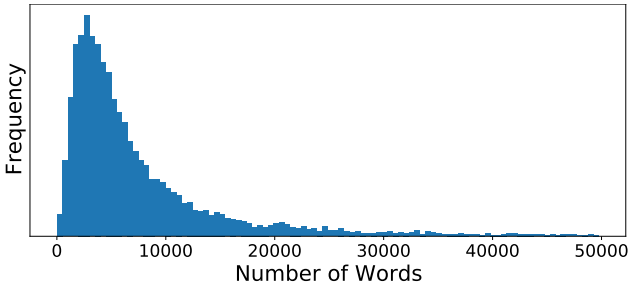


Figure 4: Number of words per opinion.

4 HIGHLIGHT EXTRACTION

In this section, we will formally define the task of highlight extraction, then describe the empirical experiments we conducted and their results.

4.1 Task Description

We formally define the task of highlight extraction as follows. Given a set of opinions V, where each opinion $v \in V$ is associated with an opinion text T_v split into a set S_v of n_v sentences. The task of highlight extraction is to predict a subset of sentences $\widehat{S_v} \subseteq S_v$ which will be cited in the future. The input of this system is the text of a single opinion. It is implied that the system's parameters are learned from the complete collection of opinions, and the system's training should not be restricted to document level tasks. Thereby, the task of identifying highlights, when formulated at sentence level can be considered as a sentence binary classification (ranking) problem.

In this work, we define this task as a sentence highlighting task. The text spans we will consider are sentences from the opinion,

we want to build model that predicts which sentence(s) in a given opinion might be cited, considering the historical observations as a ground truth that can be used for training.

This task involves the detection of network and timeline effects, as novelty arises from being the first to make a certain argument, or the first to legally qualify some facts as being compliant or not with generic statutes. As such we do expect it to be solvable at document level, at text level, if and only if we can assume that the opinion drafters are aware of the novelty of elements in the opinion, and draft them textually in a way that the task can be reduced to a Language Modeling task.

4.2 Evaluation Metrics

In line with the formulation of highlight extraction as a sentence ranking task, we will evaluate the different models with ranking metrics, such as Precision at 1 (P@1), Precision at R (P@R), Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR).

In the context of supporting the work of a legal practitioner, we value the presence of relevant sentences early in the ranked list of sentences. P@1 measures the capacity of placing a highlight sentence at the top of the ranked list, P@R is a precision metric suitable for queries with a diverse number of relevant answers, MAP balances equally important metrics precision and recall and MRR is an indicator of the effort of the system's user, in relation to the rank of the first relevant sentence.

We also consider ROUGE [25], a family of metrics for summary evaluation. For each opinion, from the list of ranked sentences, we consider the hypothesis summary to be made of the top 5 sentences. The reference summary of an opinion is made of the highlight sentences in the opinion. ROUGE will score favorably a ranker which identified sentences lexically similar to the reference sentences.

4.3 Data

We use a sample out of VerbCL Highlights as training material. We make use of a random sample of 88k opinions, from the pool of opinions that have at least one sentence marked as highlight. The data is split between training material and test material. The train collection contains circa 70k opinions.

As we have to respect the sequence length limitations of the deep neural models, we restricted the test collection to the opinions for which there is at least one highlighted sentence in the first $N = 512$ words of the opinion, and we consider the original text as made of these N first words. This test collection contains circa 1,000 opinions. Since the data is split based on opinions, the data in the test set is unseen for a model trained on the train set.

As we focus on identifying highlights, we make use of documents for which we observe that they contain highlights. We leave apart in this work the task of deciding whether or not an opinion contains highlight. From the perspective of the highlight extraction task, we consider having an oracle that has already stated the existence of highlights in the dataset we use.

4.4 Baselines

Considering the highlight task as a sentence ranking task, we propose two different types of baseline using

- binary classification at a sentence level;
- extractive summarization at a document level.

While it makes sense to consider the binary classification task, we argue that this approach will fail in this case, as most of the signals that a sentence in a legal opinion might be of interest for citation lie outside of the sentence itself. We produce this experiments to verify our assumptions that the sentence itself does not contain the signal of its importance.

An extractive summarization model will have to take the whole document into account in order to rank the sentences. On the one hand, deterministic methods will focus on one aspect for the importance of a sentence within a document. On the other hand, trainable models can be fine-tuned towards golden summaries, based on sentence-level annotations. We will use both types of models with our annotated data from VerbCL Highlights, and observe whether our annotations for importance can be learned this way. The assumption we want to confirm is that the document contains more information about the important sentence than the sentence itself, but still not enough information to lead to an accurate extraction.

We select the following models as our baselines:

- **DistilBERT** [36]: a generalized language model that we fine-tune for the binary classification task at sentence level. We use the `transformers` library from HuggingFace [10]. We derive a sentence ranker from this model by considering the estimated probability of belonging to the positive class as the sentence score.
- **TextRank** [1]: an extractive summarization deterministic algorithm. Although it can not be considered as state of the art, TextRank has proven to be a reliable baseline for summarization with a reasonable computation footprint. Considering the task of highlight detection as an extractive summarization task, we evaluate how well the standard summarization approach that ranks sentences by their importance is able to capture what is considered salient from the perspective of citing an opinion. Using the Python implementation of `PyTextRank`[30], we retrieve the score of each sentence.
- **PreSumm** [27]: an extractive summarization model, based on Bert [10]. PreSumm has the capacity to be fine-tuned on any dataset annotated at sentence level. PreSumm implementation [11] has been adapted to our dataset, and we retrieve the score of each sentence after inference. The model was trained for 50k steps, which took approximatively 14 hours with 4 nVidia GPUs.

4.5 Results

We present our results in Table 6. They confirm the initial assumptions with regard to the difficulty of the task. We formulate the highlight extraction as a task where the relevant context is the context of the current jurisprudence, therefore we make the assumption that this task can only be solved at corpus level, and not at document level. We observe from our experiences that using summarizers improves on the performance, but only to a certain degree.

The baselines we selected are introduced in their order of complexity. The sentence ranking we derive from a binary relevance

[10] https://huggingface.co/transformers/
[11] https://github.com/nlpyang/PreSumm

Model	P@1	MAP	P@R	MRR	Rouge1-F	Rouge2-F	RougeL-F
DistilBERT	0.05	0.11	0.05	0.12	0.31	0.20	0.26
TextRank	0.14	0.23	0.14	0.24	0.41	0.36	0.38
PreSumm	0.31	0.37	0.30	0.38	0.52	0.48	0.48

Table 6: Results for the Highlight Extraction Baselines

classifier based on DistilBERT performs similarly to a random ranker. This validates our assumption that highlights can not be discovered at sentence level only.

On the other hand, the improvements shown with the summarizing models emphasizes that some of the highlights can be captured at the document level. We hypothesize that this is linked to the opinion drafter's awareness of the importance of the sentence, and its potential to be cited later. An opinion drafter is a person skilled in the legal domain, they are in position to have this intuition. They would reflect the importance of the sentence at the text level, and therefore be detected by models from the BERT lineage, who excel at isolating signals from the text [9, 26, 41].

The P@1 = 0.31 observed on PreSumm is a promising result, but we also have to consider that it is restricted to short documents, while a majority of documents are longer than the maximum input length. Despite the observed improvements, we conclude that none of these systems has solved the task sufficiently.

5 RELATED WORK

In this section, we present an overview of existing work in relation with the dataset and the tasks we want to support, from within the legal domain as well as outside of this legal domain. We provide a brief overview of the previously proposed datasets, and describe the recent advancements on the tasks of highlight extraction, citation-based summarization and case law retrieval.

It is important to note that similar tasks often arise also in patent retrieval [15, 34, 37, 39, 40] and academic document retrieval domains [23, 24, 38, 44].

When drawing parallels to the tasks outside of the legal domain, the following dimensions are of a special importance:

- emergence of a directed citation network;
- time-based dependencies (it is possible to cite only previously published sources);
- highlight extraction for document summarization;
- predicting citations and text reuse.

We observe recent work in the field of citation analysis for scientific literature, enabled by S2ORC [28], which we consider a similar dataset to VerbCL Citation Graph and VerbCL Highlights in content and intent. Applied to academic literature, rating novelty [16], identifying contributions [14] or summarizing [5] share similarities with the highlight extraction task. We depart from the work done on academic literature, as we argue that the novelty is the core motivation of academic publication, while court opinions are motivated by the obligations of a public service which has to serve justice whenever it is requested. In that matter, novelty is expected to be sparser, while abundant redundancy should be the norm.

5.1 Case Law Datasets

CourtListener was also the original source of CaseLaw [29], for the purpose of testing case law retrieval models and systems, while the main contribution was a test collection of 2,500 relevance assessments and an annotation tool. We observe similarly sized datasets in other languages, such as CAIL2019-SCM [43] in Chinese. Recent meta-reviews of the field [45] provide pointers to existing datasets[12], we observe that none of them has the size required for training a complex model on a specific task.

COLIEE [18, 33] is also a relevant competition, task and dataset for legal purposes, of reduced size. It is made of multiple tasks, tackling case law retrieval with tasks 1 and 2, within the legal domain of Immigration and Citizenship in canadian courts.

Task 1 describes case law retrieval as the task of identifying which past cases should be mentioned in a query case from which citations have been removed. The rationale behind the relevance judgment, the reason why cases have been cited in the past is left unknown. Task 2 will focus on identifying relevancy at paragraph level, which is a step forward to clarifying relevance for the human reader, although the dataset does not include annotations with regard to relevance arising from either situation and facts or reasoning.

5.2 Highlight Extraction

Given a long document, the task of highlight extraction aims at finding the snippets of text that contain the information that a user will find useful in this document, making it a task ripe for user query biased extraction. For example, [19], [7] and [32] show the importance of highlighting for learning. Highlighting is described as a manual annotation task, for the benefit of meeting future information needs. The concept is put in practice also by [13] and [8], where highlights support the generation or the evaluation of text summaries. We observe only few attempts at creating an automated highlighting system for text, in contrast to the flourishing field of video highlight generation.

The key concept we articulate our task around is meeting future information needs. Although we have plenty of historical data available, we can only speculate now on what is important in any opinion that is published, and what will be picked up in cases that courts will litigate in the near or far future.

The highlight extraction task we present in this work focus on identifying the novel legal content in a new case that would potentially be cited in future cases, and evaluate how this could stem from text level analysis of the opinion, rather than identifying how an opinion would suit an argument made in another one. We refer to the Section 4 for more details.

[12] https://github.com/thunlp/LegalPapers

5.3 Case Law Retrieval

Our work is also related to the previous work on the task of case law retrieval. In this task, a legal practitioner retrieves a list of past cases that are legally relevant to a case at hand.

COLIEE [18, 33] has hosted a yearly competition including such a retrieval task, restricted to cases from the Immigration and Citizenship Courts of Canada. The target is to identify cases that could be cited in a query case.

The task is reduced to identifying the similarities between court opinions, similarities in details of the case, in the underlying storylines and reasoning. Participants reformulated the retrieval task as a ranking task based on a pairwise relevancy classifier, with various text representation techniques applied, from lexical representations [20, 42] to deep representations [11, 35]. It is observed that the scarcity of data is hindering the use of advanced data-driven models. This domain relevant task has shown the difficulty to work at text level to evaluate a relevancy that professionals had annotated.

We see a potential application of our dataset in a case law retrieval task guided by arguments, where the query is an argument, and the search results a list of past opinions that could support the argument, based on the highlights that were extracted.

5.4 Citation-based Summarization

The task of highlight extraction finds its roots in the task of summarizing long documents. We observed many efforts in the field of summarizing by using citations, in the domain of academic literature summarization, for example. Citation-based summarization was first introduced in the context of scholarly data processing for summarizing scientific articles [17]. In citation-based summarization, citation anchors (called "catchphrases") for one target document over multiple citing document form a summary for this target document. The CL-SciSumm shared task [6] aims at producing summaries of academic papers guided by citation anchors (called "citances"), by identifying spans of text relevant to the anchor. The approaches are evaluated against a set of golden summaries, using standard metrics of the ROUGE family [25].

The court opinions will show a predetermined structure, as academic papers do, although the articulation in different parts or chapters will not be signaled explicitly in the text. Our task definition differs as we focus on actual verbatim quotes from cited documents which are less biased towards the writer's view of the cited document than paraphrases. Our work considers these verbatim quotes to be the target summary, as opposed to considering them as a way to build a summary similar to the golden summary.

We consider that the spans of an opinion which are later used in citations form a summary of this opinion, which a specific type of summary guided by the search for novel and important legal aspects addressed in the opinion document. This summary characterizes what the change an opinion makes in terms of the legal argumentation, i.e., its network effect (global context), not just a summary of all the elements present in the opinion (local context).

6 CONCLUSION

In this paper we introduced the task of highlight extraction from court opinions and the large dataset VerbCL of annotated court opinions aimed at supporting training and evaluation for the highlight extraction task. Our dataset focuses on citations made in a court opinions by quoting verbatim preceding court opinions in support of a legal argument. VerbCL is sufficiently large and can be used for training machine learning models on the highlight extraction task. We also see the potential of VerbCL for other legal information retrieval tasks that can be informed by the citation network.

We also demonstrated the difficulty of the highlight extraction task, which escapes the reduction to a sentence-level or document-level task. Note that the citation network is the result of a massively distributed expert work from the date of the first citation until now. Every citation is the result of an expert's retrieval of a relevant opinion to support the argument in the context of a specific case at hand. However, we can verify the importance of every specific text span from an opinion only post-hoc, i.e., not at the time of the publication but later on as we observe how the same arguments made in the opinion are subsequently re-used within other opinion documents in our dataset. Only this reuse over time indicates the document's importance in terms of the impact it made to help shape other documents. Future work should aim to address such network and historical effects, distilling knowledge of the previously published documents into the processing of a new one. We believe that similar approaches will also prove useful for patent retrieval and academic search scenarios.

Future work should also consider extending the analysis to abstractive anchors, which requires a text generation task setup instead of purely retrieval-based one considered here. Finally, considering the full document length remains a challenge for the current neural ranking approaches that should be addressed to be applicable in the legal domain.

ACKNOWLEDGMENTS

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), the Amsterdam Business School PhD program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the Similarity Function of TextRank for Automated Summarization. arXiv:1602.03606 [cs.CL]

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[3] Ulrik Brandes. 2001. A faster algorithm for betweenness centrality. *Journal of mathematical sociology* 25, 2 (2001), 163–177.

[4] Ulrik Brandes and Christian Pich. 2007. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos* 17, 07 (2007), 2303–2318.

[5] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011* (2020).

[6] Muthu Kumar Chandrasekaran, Michihiro Yasunaga, Dragomir Radev, Dayne Freitag, and Min-Yen Kan. 2019. Overview and Results: CL-SciSumm Shared Task 2019. *CEUR Workshop Proceedings* 2414 (jul 2019), 153–166. arXiv:1907.09854 http://arxiv.org/abs/1907.09854

[7] X. Che, H. Yang, and C. Meinel. 2018. Automatic Online Lecture Highlighting Based on Multimedia Analysis. *IEEE Transactions on Learning Technologies* 11, 1 (2018), 27–40. https://doi.org/10.1109/TLT.2017.2716372

[8] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany, 484–494. https://doi.org/10.18653/v1/P16-1046

[9] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. arXiv:1906.04341 [cs.CL]

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Baban Gain, Dibyanayan Bandyopadhyay, Tanik Saikh, and Asif Ekbal. 2019. IITP in COLIEE@ICAIL 2019: Legal Information Retrieval using BM25 and BERT. (06 2019). https://doi.org/10.13140/RG.2.2.28887.32161

[12] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, Gaël Varoquaux, Travis Vaught, and Jarrod Millman (Eds.). Pasadena, CA USA, 11 – 15.

[13] Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based Reference-less Evaluation of Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Florence, Italy, 3381–3392. https://doi.org/10.18653/v1/P19-1330

[14] Hiroaki Hayashi, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. What's New? Summarizing Contributions in Scientific Literature. arXiv:2011.03161 [cs.CL]

[15] Sebastian Hofstätter, Navid Rekabsaz, Mihai Lupu, Carsten Eickhoff, and Allan Hanbury. 2019. Enriching word embeddings for patent retrieval with global context. In *European Conference on Information Retrieval.* Springer, 810–818.

[16] Bolin Hua and YoungKug Shin. 2021. Extraction of Sentences Describing Originality from Conclusion in Academic Papers. (2021).

[17] Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2020. Automatic summarization of scientific articles: A survey. https://doi.org/10.1016/j.jksuci.2020.04.020

[18] Yoshinobu Kano, Miyoung Kim, M. Yoshioka, Yao Lu, J. Rabelo, Naoki Kiyota, R. Goebel, and K. Satoh. 2018. COLIEE-2018: Evaluation of the Competition on Legal Information Extraction and Entailment. In *JSAI-isAI Workshops.*

[19] David Y. J. Kim, Adam Winchell, Andrew E. Waters, Phillip J. Grimaldi, Richard G. Baraniuk, and Michael C. Mozer. 2020. Inferring student comprehension from highlighting patterns in digital textbooks: An exploration of an authentic learning platform. In *Proceedings of the Second International Workshop on Intelligent Textbooks.*

[20] Moemedi Lefoane, Tshepho Koboyatshwene, and Lakshmi Narasimhan. 2018. KNN clustering approach to legal precedence retrieval.

[21] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (Feb. 1966), 707.

[22] Loet Leydesdorff, Caroline S Wagner, and Lutz Bornmann. 2018. Betweenness and diversity in journal citation networks as measures of interdisciplinarity—A tribute to Eugene Garfield. *Scientometrics* 114, 2 (2018), 567–592.

[23] Xinyi Li and Maarten de Rijke. 2017. Do topic shift and query reformulation patterns correlate in academic search?. In *European Conference on Information Retrieval.* Springer, 146–159.

[24] Xinyi Li, Bob JA Schijvenaars, and Maarten de Rijke. 2017. Investigating queries and search failures in academic search. *Information processing & management* 53, 3 (2017), 666–683.

[25] Chin-Yew Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries.* Technical Report. 74–81 pages. https://www.aclweb.org/anthology/W04-1013

[26] Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* Association for Computational Linguistics, Florence, Italy, 241–253. https://doi.org/10.18653/v1/W19-4825

[27] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. *CoRR* abs/1908.08345 (2019). arXiv:1908.08345 http://arxiv.org/abs/1908.08345

[28] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[29] Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018.* Association for Computing Machinery, Inc, New York, NY, USA, 1261–1264. https://doi.org/10.1145/3209978.3210161

[30] Paco Nathan. 2016. *PyTextRank, a Python implementation of TextRank for phrase extraction and summarization of text documents.* https://doi.org/10.5281/zenodo.4637885

[31] José Luis Ortega. 2014. Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search. *Journal of Informetrics* 8, 3 (2014), 728–737. https://doi.org/10.1016/j.joi.2014.07.001

[32] Héctor R. Ponce, Richard E. Mayer, Verónica A. Figueroa, and Mario J. López. 2018. Interactive highlighting for just-in-time formative assessment during whole-class instruction: effects on vocabulary learning and reading comprehension. *Interactive Learning Environments* 26, 1 (2018), 42–60. https://doi.org/10.1080/10494820.2017.1282878 arXiv:https://doi.org/10.1080/10494820.2017.1282878

[33] Juliano Rabelo, mi-young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. *A Summary of the COLIEE 2019 Competition.* 34–49. https://doi.org/10.1007/978-3-030-58790-1_3

[34] Julien Rossi and Evangelos Kanoulas. 2018. Query Generation for Patent Retrieval with Keyword Extraction Based on Syntactic Features.. In *JURIX.* 210–214.

[35] Julien Rossi and Evangelos Kanoulas. 2019. Legal Information Retrieval with Generalized Language Models. *Proceedings of the 6th Competition on Legal Information Extraction/Entailment. COLIEE* (2019).

[36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL]

[37] Serhad Sarica, Binyang Song, En Low, and Jianxi Luo. 2019. Engineering knowledge graph for keyword discovery in patent search. In *Proceedings of the Design Society: International Conference on Engineering Design*, Vol. 1. Cambridge University Press, 2249–2258.

[38] Philipp Schaer, Johann Schaible, and Leyla Jael Garcia Castro. 2020. Overview of LiLAS 2020–Living Labs for Academic Search. In *International Conference of the Cross-Language Evaluation Forum for European Languages.* Springer, 364–371.

[39] Walid Shalaby and Wlodek Zadrozny. 2018. Toward an interactive patent retrieval framework based on distributed representations. In *The 41st International ACM SIGIR conference on research & development in information retrieval.* 957–960.

[40] Walid Shalaby and Wlodek Zadrozny. 2019. Patent retrieval: a literature review. *Knowledge and Information Systems* (2019), 1–30.

[41] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. arXiv:1905.05950 [cs.CL]

[42] Vu Tran, Son Truong Nguyen, and Minh Le Nguyen. 2018. JNLP group: legal information retrieval with summary and logical structure analysis. In *Twelfth International Workshop on Juris-informatics (JURISIN), COLLEE.*

[43] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. CAIL2019-SCM: A Dataset of Similar Case Matching in Legal Domain. *CoRR* abs/1911.08962 (2019). arXiv:1911.08962 http://arxiv.org/abs/1911.08962

[44] Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web.* 1271–1279.

[45] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 5218–5230. https://doi.org/10.18653/v1/2020.acl-main.466