Measuring Semantic Coherence of a Conversation

Svitlana Vakulenko, Maarten de Rijke,
Michael Cochez, Vadim Savenkov, Axel Polleres

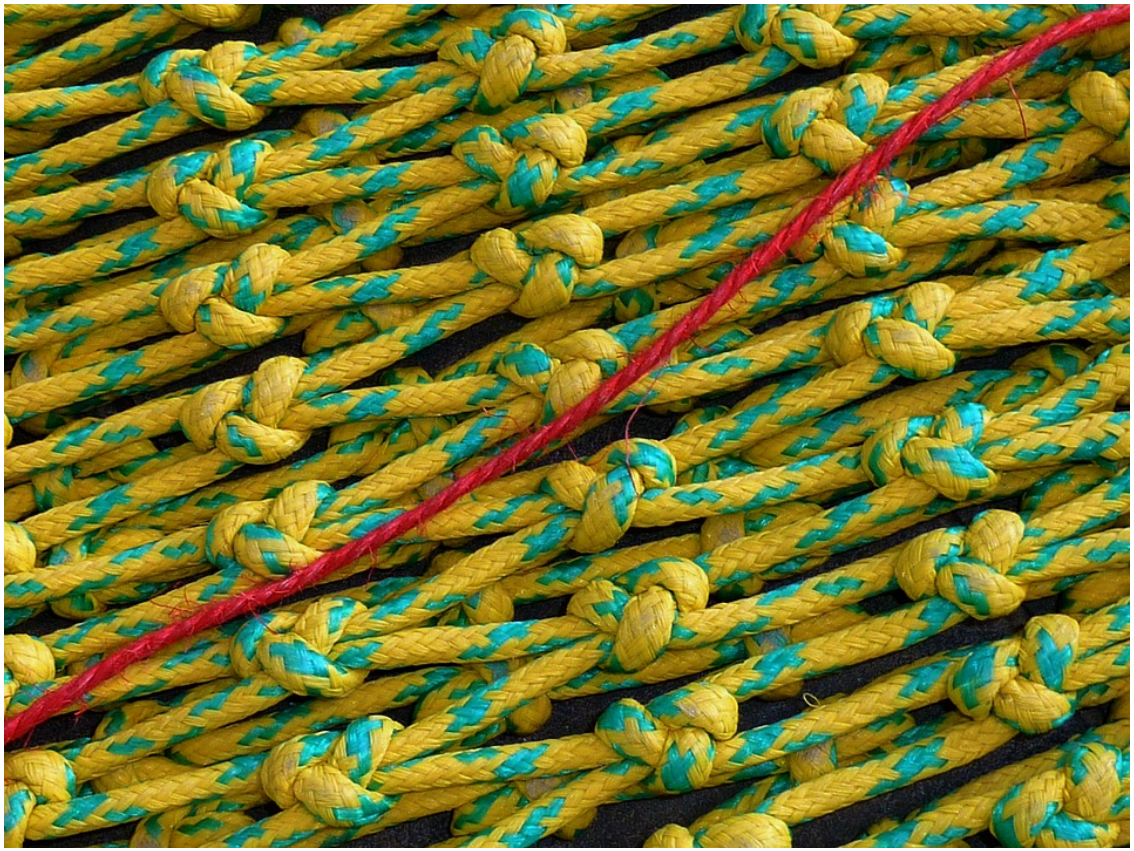
WU
WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

6 JUNE 2018

Semantic coherence

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

- An essential property of a conversation, "**continuity** of senses"



https://pixabay.com/en/fishing-net-red-thread-network-node-1526496/

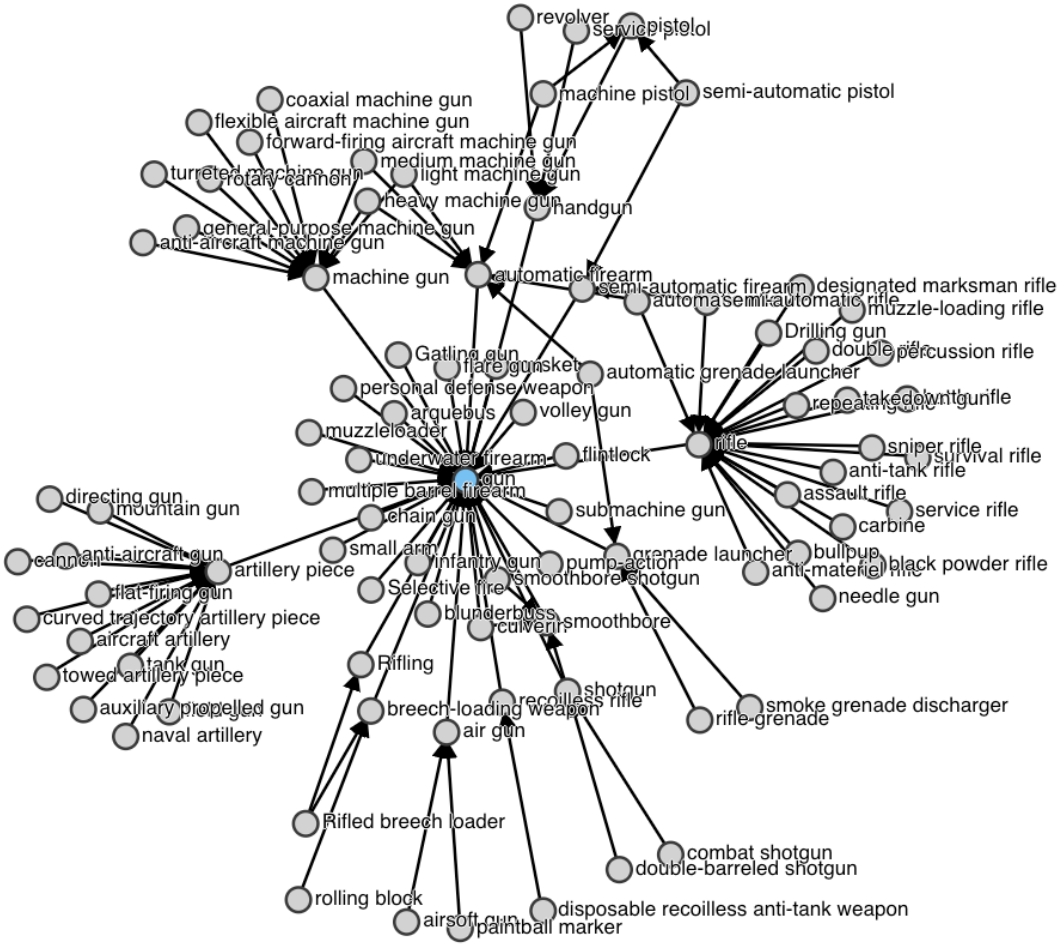
Research goal

- See if we can **detect** holes in conversations
- Evaluate existing **knowledge** models
- Propose an approach to **measure** these holes (incoherence)

- **Why**: dialogue system design, knowledge engineering

Semantic models

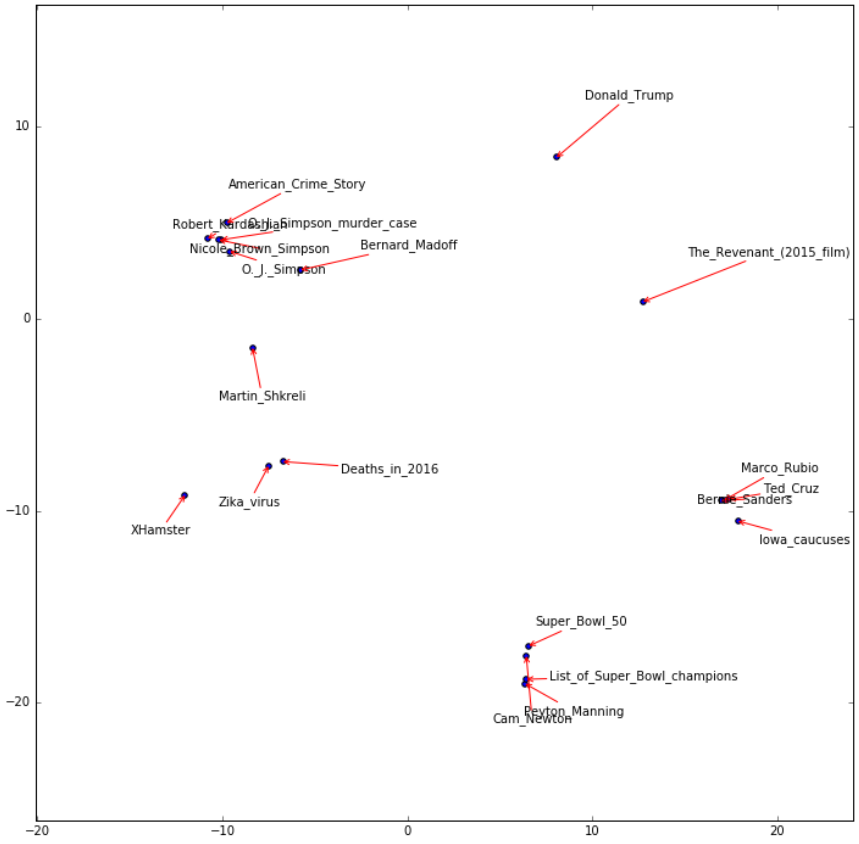
- Knowledge Graphs



https://commons.wikimedia.org/wiki/File:Wikidata-gun-ontology-2017-05-11.png

Semantic models

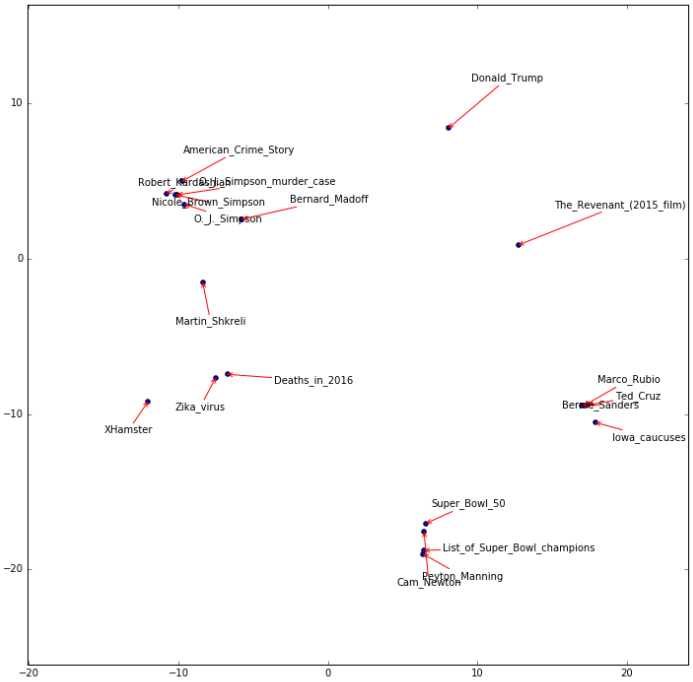
- Word embeddings

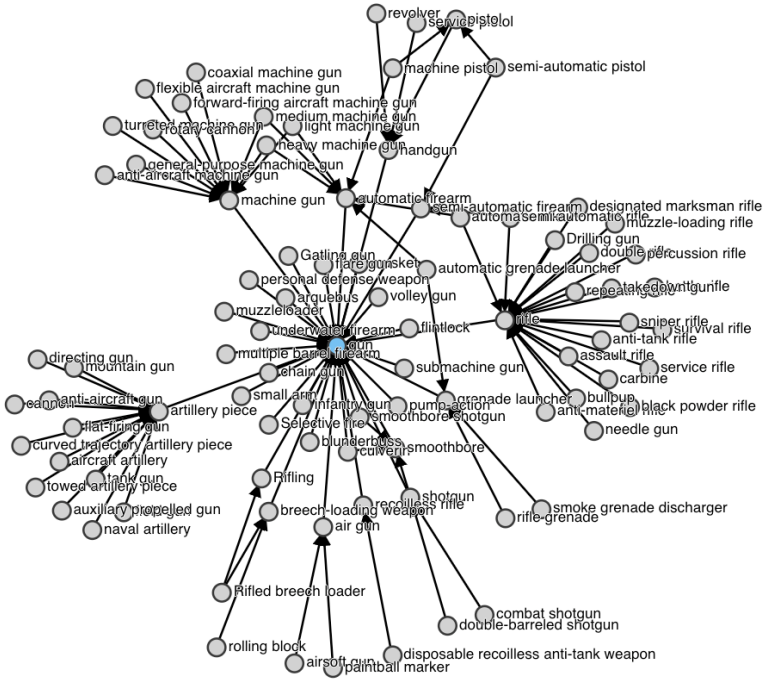


https://commons.wikimedia.org/wiki/File:2016_02_mini_embedding.png

Semantic models

- Knowledge Graphs
- Word embeddings



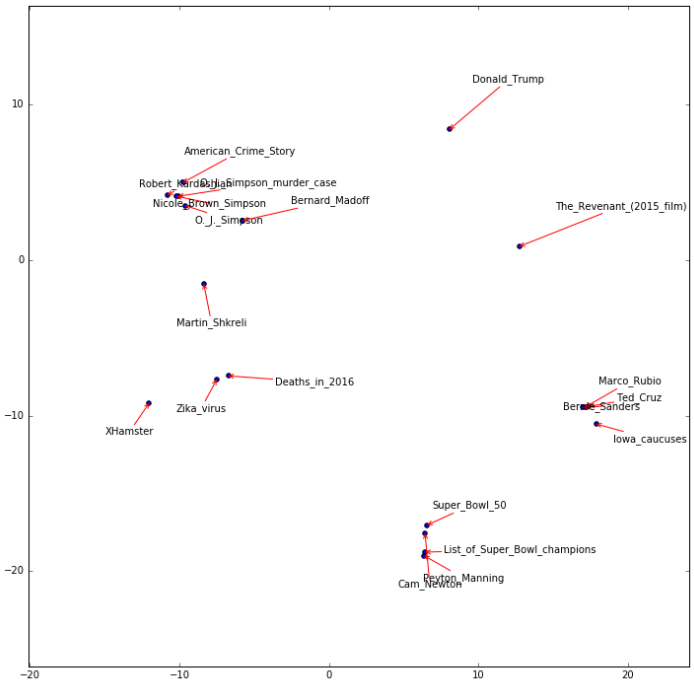


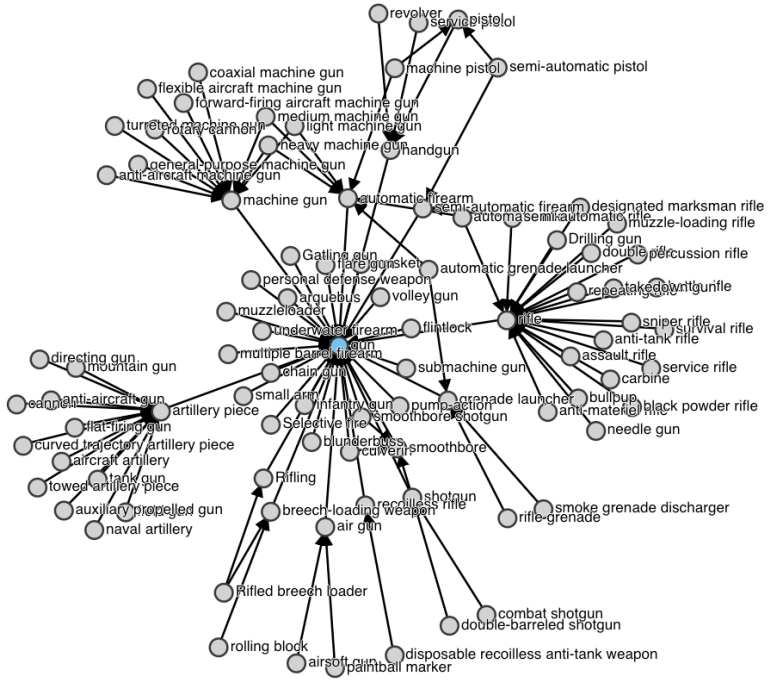
https://commons.wikimedia.org/wiki/File:2016_02_mini_embedding.png

https://commons.wikimedia.org/wiki/File:Wikidata-gun-ontology-2017-05-11.png

Semantic models

- Knowledge Graphs
- Word embeddings
- Knowledge Graph embeddings



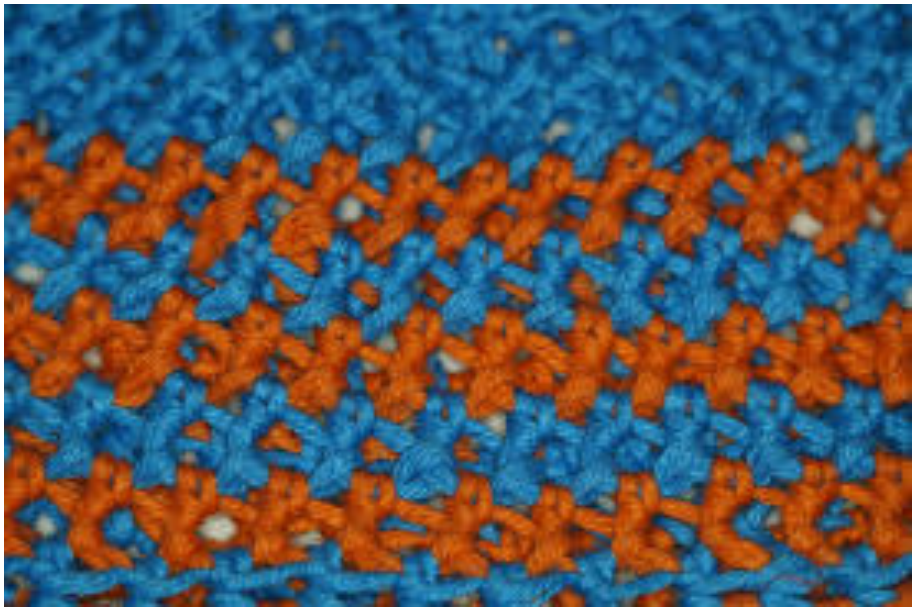


https://commons.wikimedia.org/wiki/File:2016_02_mini_embedding.png

https://commons.wikimedia.org/wiki/File:Wikidata-gun-ontology-2017-05-11.png

Linking dialogue

- Take existing **knowledge** models
- See if we can **detect** holes in conversations through this models
- Propose an approach to **measure** these holes (incoherence)



https://pxhere.com/en/photo/1101883

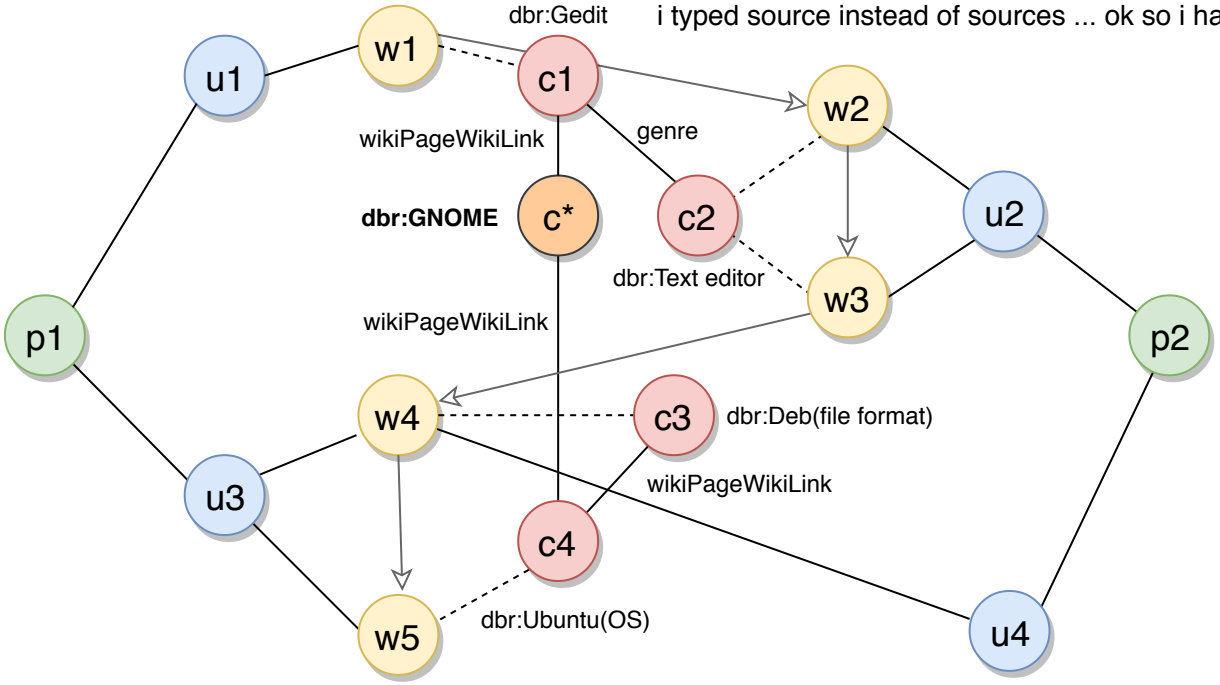
Dialog graph

w1

mdg: gksudo **gedit** /etc/apt/source.list
(type from command line)

w2 w3

crunchbang666: the **text editor** has opened the file
source.list but there is no content
i typed source instead of sources ... ok so i have it open



w4 w5

mdg: see the line # **deb** http://gb.archive.**ubuntu**
all you have to do is delete the ""#"" character

w4

crunchbang666: just the **deb** or the deb-src line too?

Experiments

- Ubuntu Dialogue Corpus
- DBpedia Spotlight API
- Knowledge Graphs: DBpedia+Wikidata HDT
- Knowledge Graph embeddings: rdf2vec, KGlove
- Word embeddings: word2vec, Glove

https://github.com/rkadlec/ubuntu-ranking-dataset-creator

https://en.wikipedia.org/wiki/File:DBpediaSpotlight.jpg

https://en.wikipedia.org/wiki/Wikidata

Subgraph induction

top-k shortest path

```
PREFIX ppf: <java:at.ac.wu.arqext.path.>
PREFIX dbr: <http://dbpedia.org/resource/>
SELECT * WHERE {
?X ppf:topk ("--source" dbr:Directory_service
                    dbr:Gnome dbr:GNOME
                    dbr:Desktop_environment
                "--target" dbr:Desktop_computer
                "--k" 5 "--maxlength" 9 "--timeout" 2000) }
```

http://wikidata.communidata.at

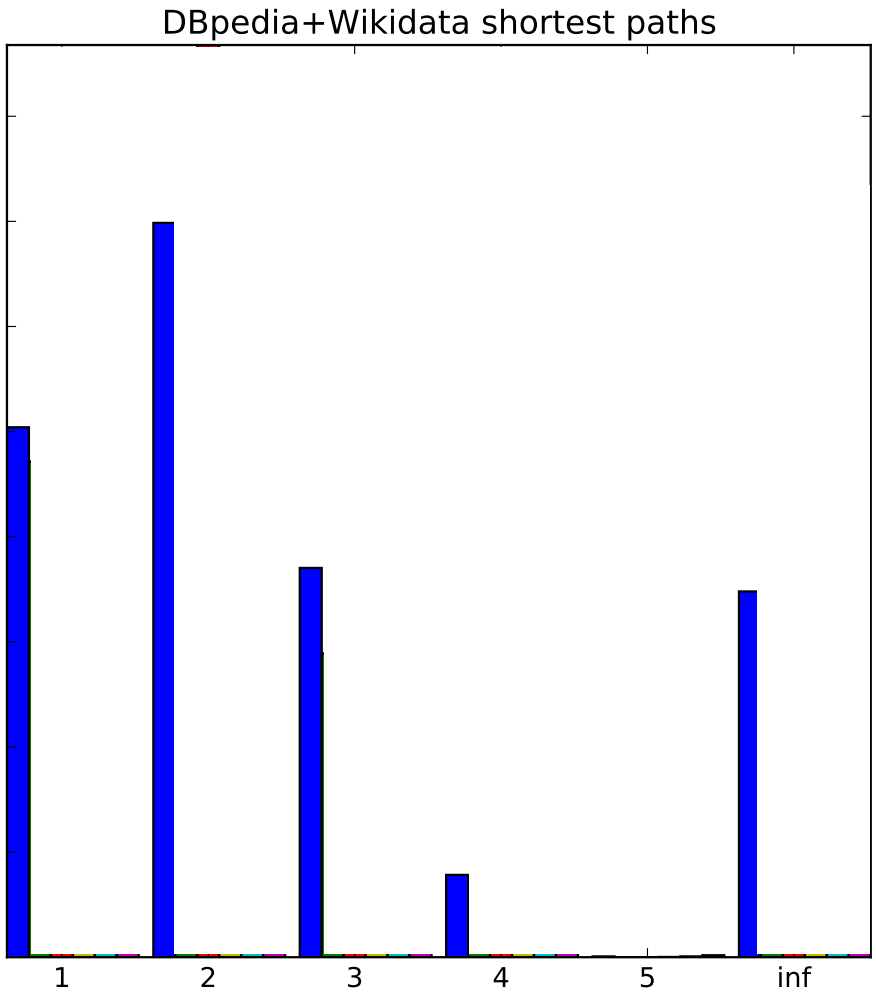
Subgraph statistics

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Table 1. The top 5 most common entities and relations in the Ubuntu dialogue dataset: mentioned entities – from linking dialogue utterances to DBpedia entities via Dbpedia Spotlight Web service; context entities and relations – from the shortest paths between the mentioned entities in DBpedia.

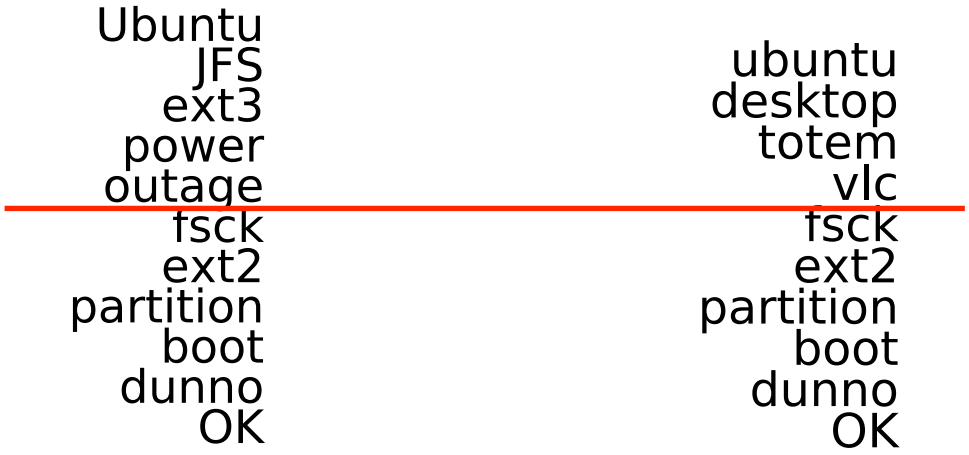
Top	Mentioned entities		Context entities		Relations	
#	Label	Count	Label	Count	Label	Count
1	Ubuntu(philosophy)	1605	Ubuntu(OS)	1058	wikiPageWikiLink	51014
2	Sudo	708	Linux	725	gold/hypernym	319
3	Booting	676	Microsoft_Windows	208	ontology/genre	178
4	APT(Debian)	405	FreeBSD	175	operatingSystem	140
5	Live_CD	314	Smartphone	171	rdf-schema#seeAlso	116

Shortest paths

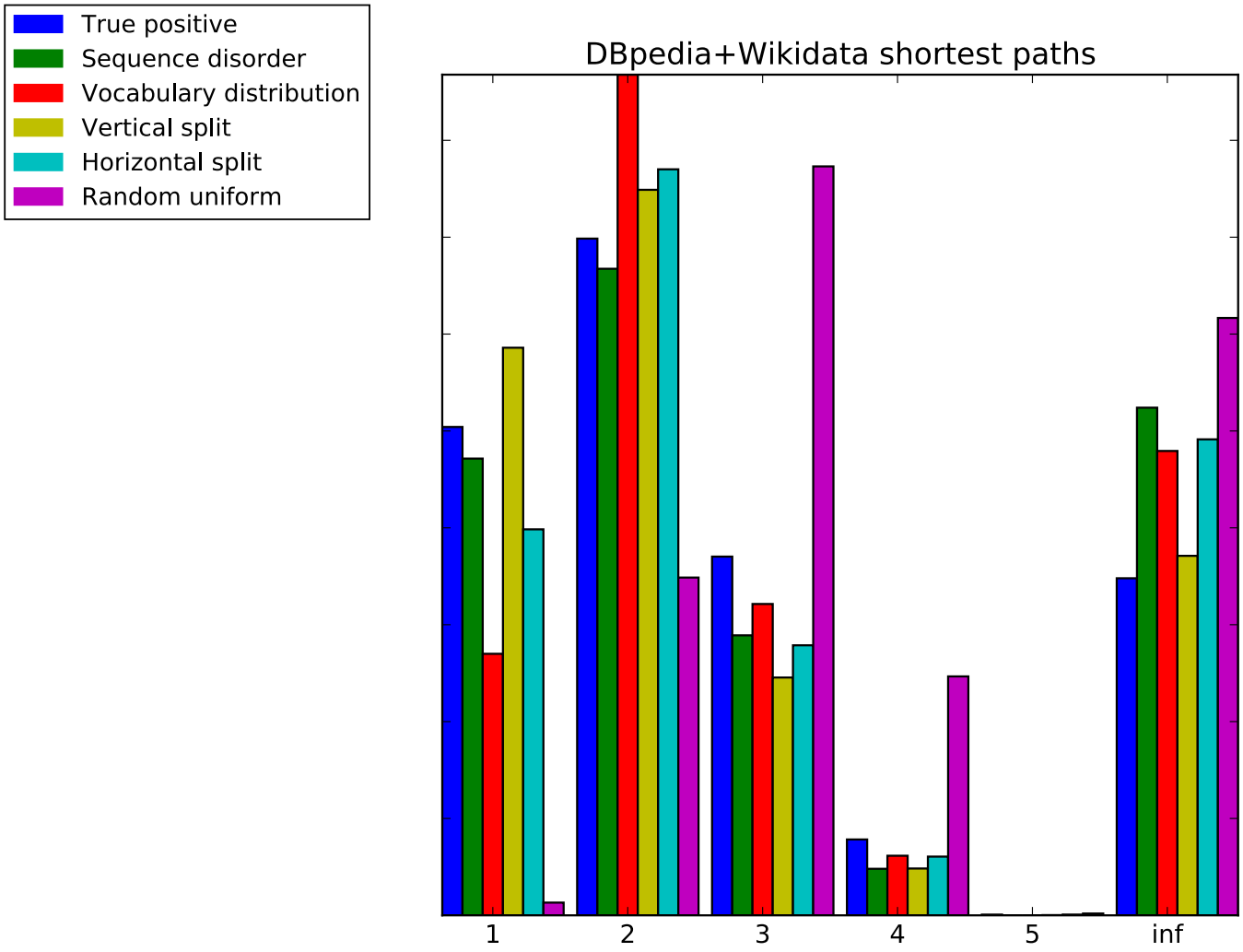


Negative sampling

- random uniform (RUf)
- vocabulary distribution (VoD)
- sequence disorder (SqD)
- horizontal split (HSp)
- vertical split (VSp)

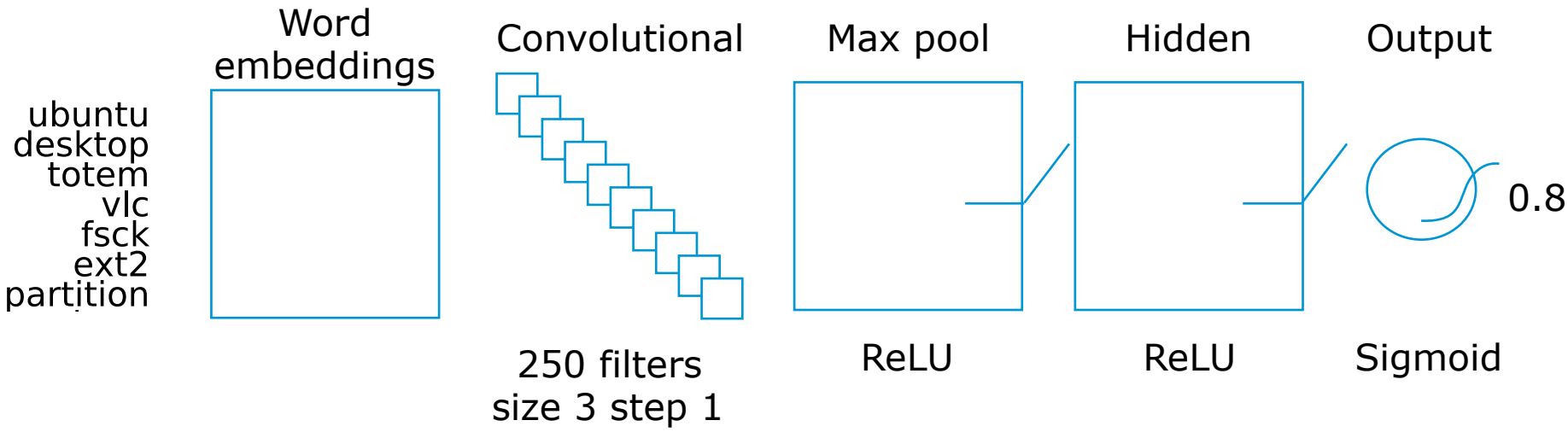


Shortest paths



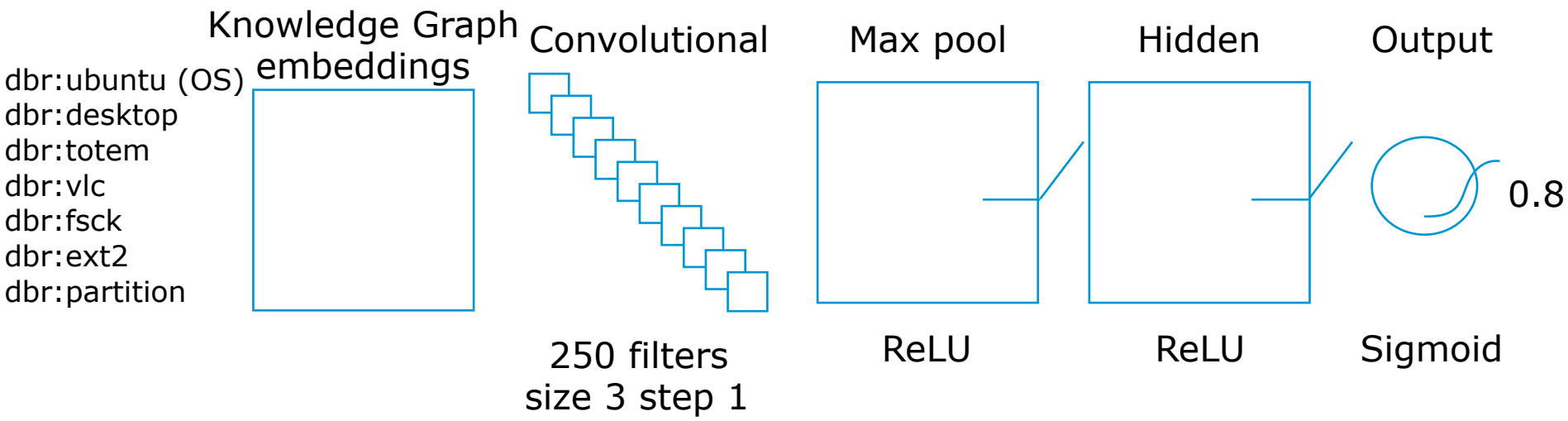
Binary classification

- Convolutional Neural Network (CNN)
- Input: sequence of words/entities
- Output: coherence score [0;1]



Binary classification

- Convolutional Neural Network (CNN)
- Input: sequence of words/entities
- Output: coherence score [0;1]



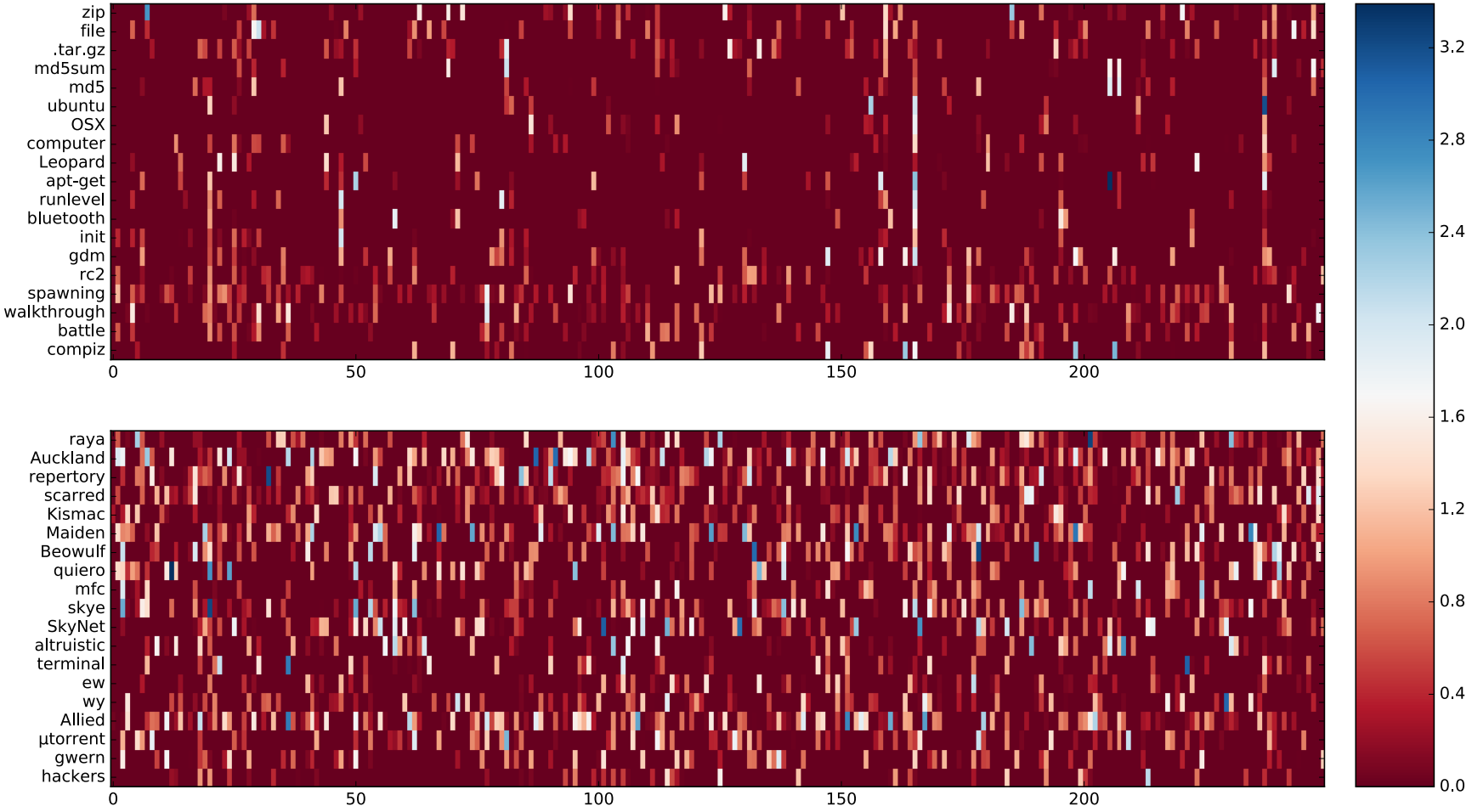
Results

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

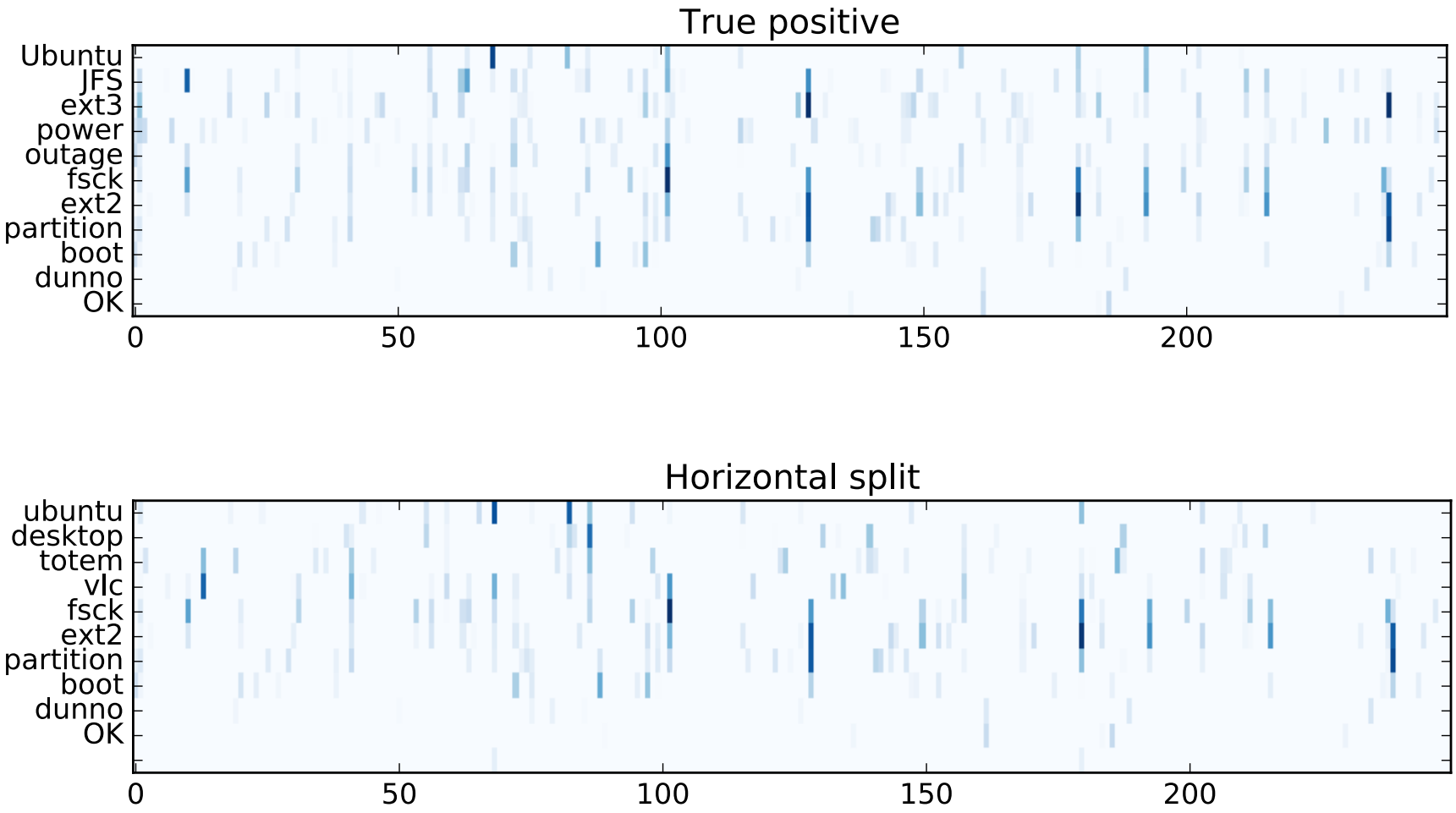
Embeddings	Data split	Accuracy											
		TPos	TNeg										Avg
			RUf	Avg	VoD	Avg	SqD	Avg	VSp	Avg	HSp	Avg	
Word2Vec	RUf	0.99	0.99	**0.99**	0.02	0.50	0.02	0.50	0.01	0.50	0.01	0.50	0.60
	VoD	0.89	0.62	0.75	0.90	0.89	0.53	0.71	0.18	0.54	0.20	0.54	**0.69**
	SqD	0.75	0.65	0.70	0.88	0.81	0.81	0.78	0.27	0.51	0.29	0.52	0.66
	VSp	0.59	0.50	0.55	0.82	0.71	0.41	0.50	0.59	0.59	0.61	0.60	0.59
	HSp	0.62	0.39	0.50	0.71	0.66	0.38	0.50	0.55	0.58	0.63	0.63	0.58
GloVe	RUf	0.99	0.99	**0.99**	0.00	0.50	0.01	0.50	0.00	0.50	0.00	0.50	0.60
	VoD	0.93	0.38	0.66	0.93	**0.93**	0.39	0.66	0.19	0.56	0.08	0.51	0.66
	SqD	0.76	0.71	0.73	0.91	0.84	0.82	**0.79**	0.16	0.46	0.15	0.45	0.66
	VSp	0.60	0.25	0.42	0.92	0.76	0.43	0.51	0.65	0.62	0.66	0.63	0.59
	HSp	0.71	0.34	0.52	0.81	0.76	0.30	0.50	0.55	**0.63**	0.66	**0.68**	0.62
rdf2vec PRS	RUf	0.98	0.99	**0.99**	0.02	0.50	0.02	0.50	0.02	0.50	0.01	0.50	0.60
	VoD	0.79	0.68	0.73	0.83	0.81	0.34	0.57	0.36	0.57	0.35	0.57	0.65
	SqD	0.59	0.48	0.54	0.72	0.66	0.67	0.63	0.43	0.51	0.40	0.50	0.56
rdf2vec PR	HSp	0.57	0.59	0.58	0.72	0.64	0.43	0.50	0.59	0.58	0.67	0.62	0.58
KGloVe Uni	RUf	0.92	0.97	0.94	0.11	0.51	0.09	0.50	0.08	0.50	0.07	0.50	0.59
	VoD	0.54	0.88	0.71	0.73	0.64	0.61	0.58	0.51	0.52	0.52	0.53	0.60
	SqD	0.55	0.62	0.58	0.64	0.59	0.63	0.59	0.47	0.51	0.45	0.50	0.56
KGloVe PrO	HSp	0.31	0.81	0.56	0.75	0.53	0.69	0.50	0.77	0.54	0.70	0.51	0.53
KGloVe PR	HSp	0.47	0.69	0.58	0.61	0.54	0.54	0.50	0.57	0.52	0.65	0.56	0.54

Random uniform

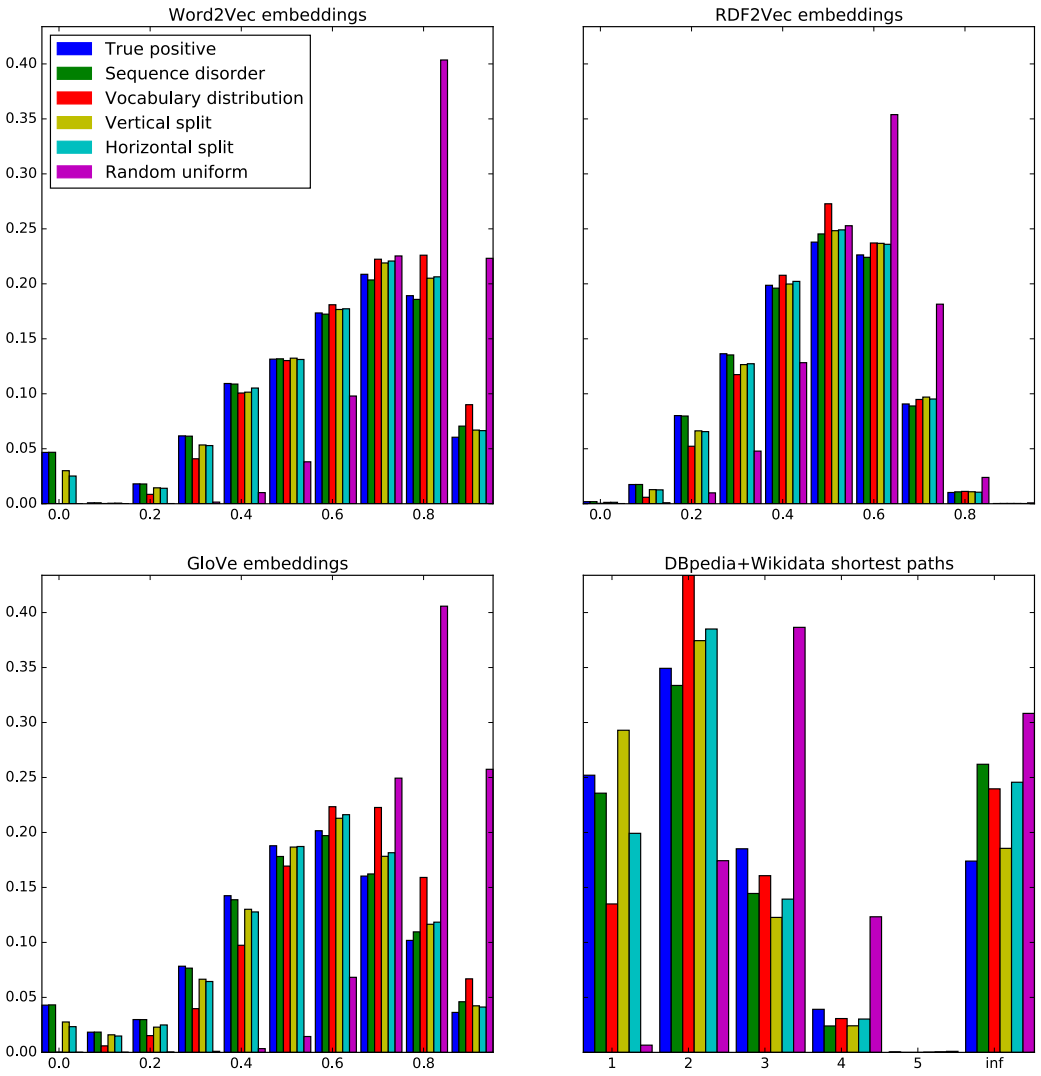
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS



Horizontal split

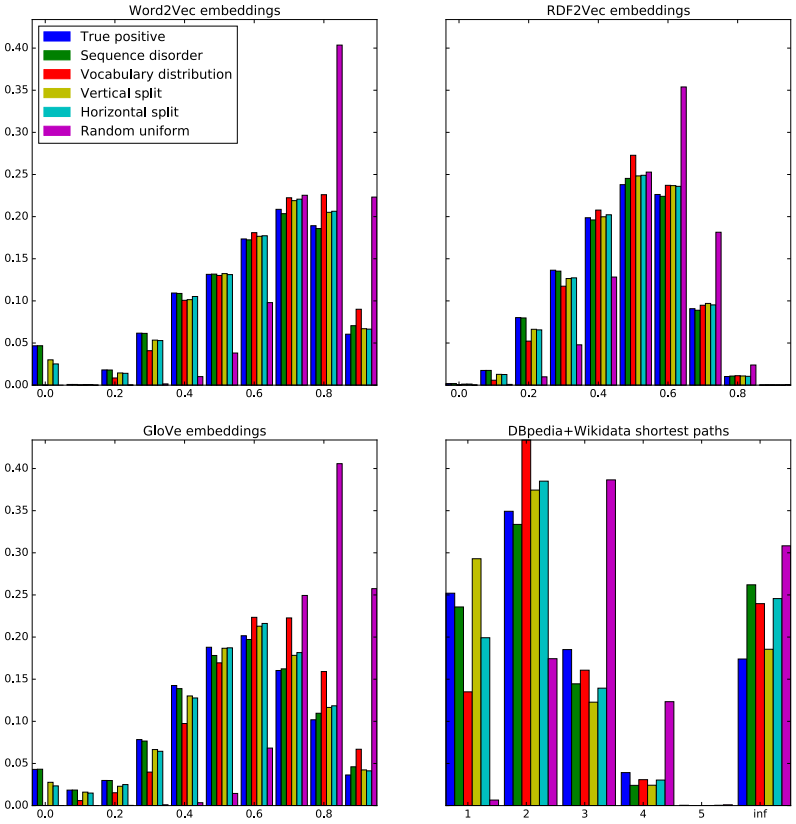


Semantic spaces



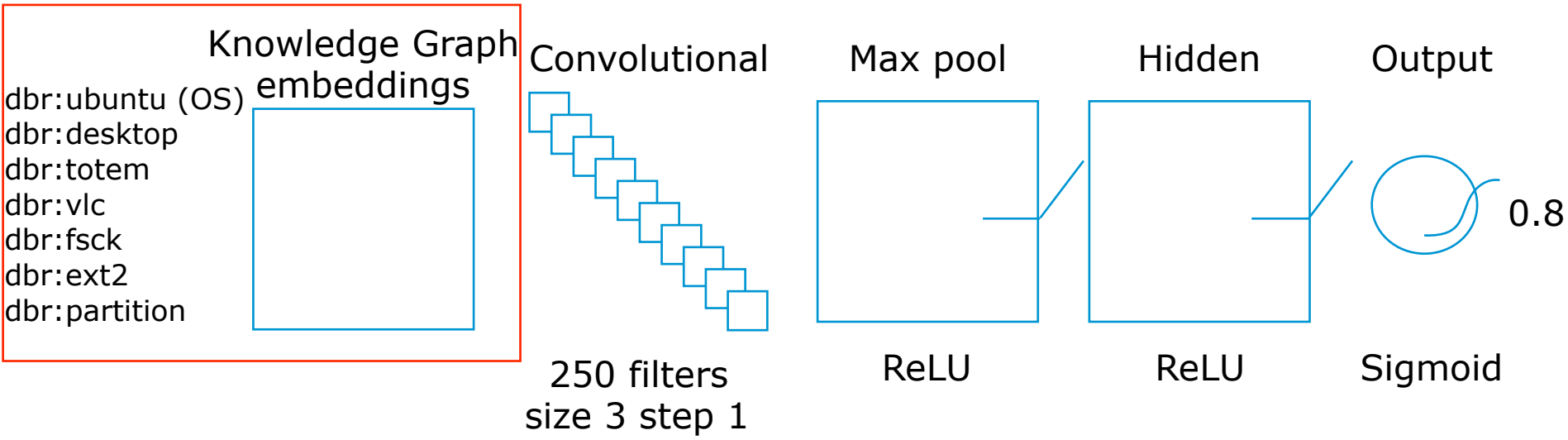
Conclusions and future work

- GloVe **word embeddings** show best performance
  - integrating **heterogenous** knowledge sources



Conclusions and future work

- **NEL** is a bottleneck for KG embeddings
  - **End-to-end** training (NEL NN-layer)



Conclusions and future work

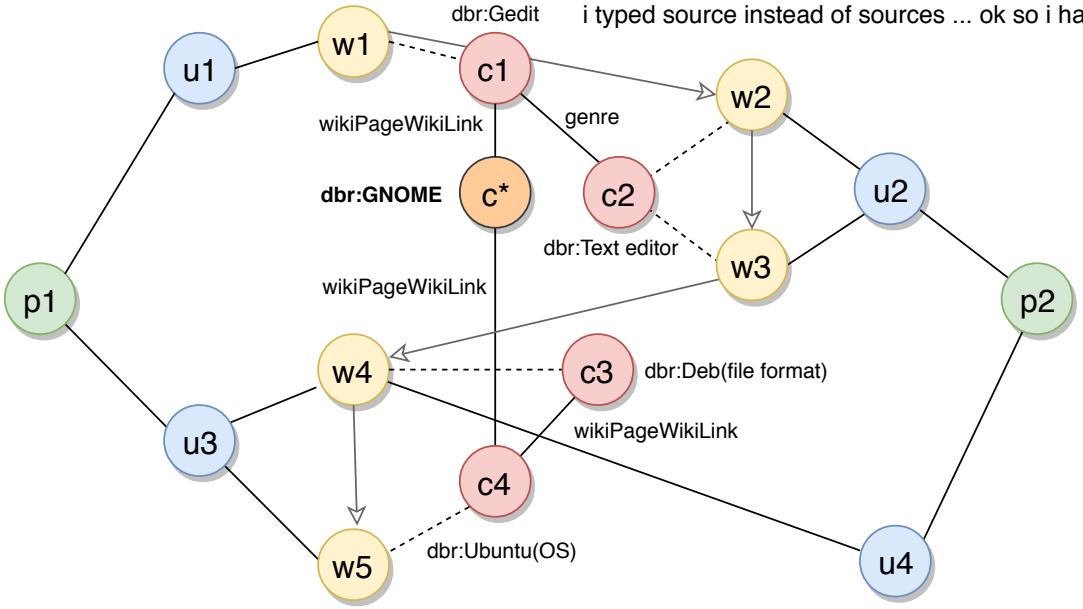
- **Dialog graph** embeddings

w1

mdg: gksudo **gedit** /etc/apt/source.list
(type from command line)

w2 w3

crunchbang666: the **text editor** has opened the file source.list but there is no content
i typed source instead of sources ... ok so i have it open



w4 w5

mdg: see the line # **deb** http://gb.archive.**ubuntu**
all you have to do is delete the ""#"" character

w4

crunchbang666: just the **deb** or the deb-src line too?