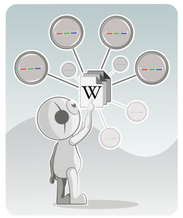


OPIEC: An Open Information Extraction Corpus

Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, Rainer Gemulla

University of Mannheim, Mannheim, Gemany

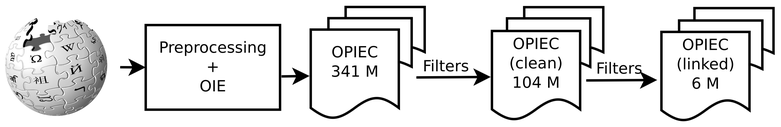
UNIVERSITY OF MANNHEIM

OPIEC: Open Information Extraction Corpus

The **largest OIE corpus** to date

1. Aims to spur research in **AKBC**, **open Q&A**, ...
2. **Rich with meta-data** – many syntactic/semantic annotations
3. Multiple **sub-corpora** from noisy to clean
4. Analyzed and compared with **Wikipedia-based KBs**

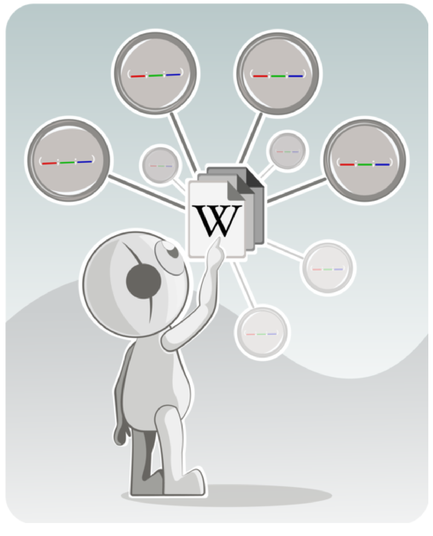
Corpus Construction



Bonus corpus: WikiNLP - rich linguistic annotations for every article of Wikipedia

- depdenency parse, POS tags, NER tags, spans, ...

Resources

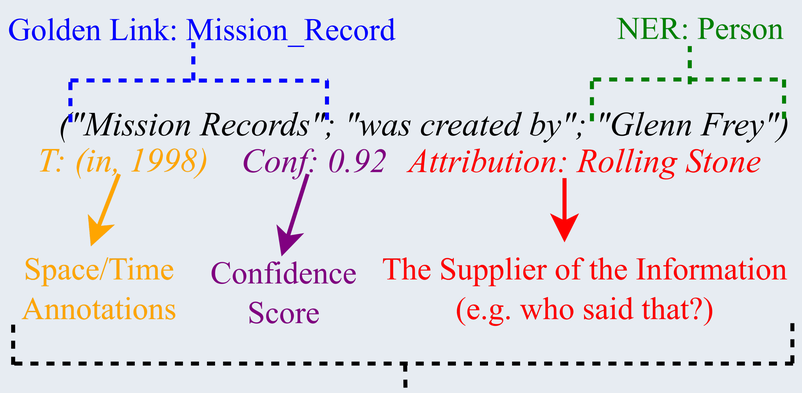


https://www.uni-mannheim.de/dws/research/resources/opiec

OPIEC-Raw (341M Triples)

Constructed using the state-of-the-art OIE system **MinIE-SpaTe** on Wikipedia

Example Triple:



Applications: such corpora are valuable resources for downstream tasks

- Automated KB construction, open question answering, event schema induction, etc.

OPIEC-Link (∼6M Triples)

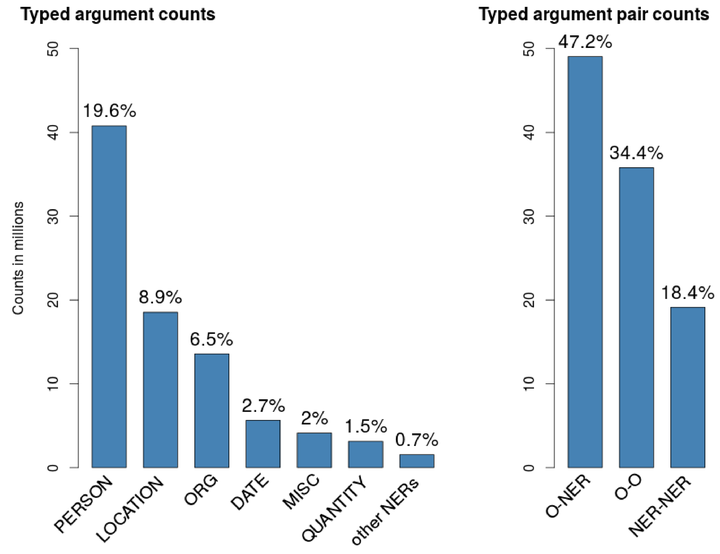
Subcorpus of **linked extraction**

- Linked = both arguments are linked to **entities** and **concepts**
- The original **golden links** from Wikipedia articles are kept
- The **largest corpus** to date with **golden disambiguation links** for the arguments
- Facilitates corpus analysis

OPIEC-Clean (104M Triples)

Subcorpus of **clean extractions**

- Clean = arguments are **self-contained** and refer to **concepts**
- Higher confidence, generally shorter



Analysis: OPIEC and Knowledge Bases

Goal: compare OPIEC triples to KBs

- Triple is **KB hit** when potentially present in KB (optimistically measured)

Example: The most frequent open relations aligned to DBpedia relations

associatedMusicalArtist		spouse	
"be"	(5,521)	*"be wife of"*	(1,580)
"have"	(3,248)	*"be"*	(980)
"be guitarist of"	(619)	*"marry"*	(551)
"be drummer of"	(433)	*"be widow of"*	(392)
"be feature"	(377)	*"be marry to"*	(246)
"be frontman of"	(367)	*"have"*	(244)