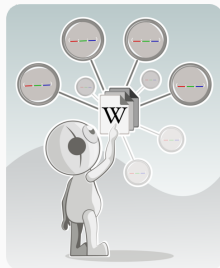


OPIEC: An Open Information Extraction Corpus

Kiril Gashteovski

University of Mannheim
Data and Web Science Group



Open Information Extraction (OIE)

- **Goal:** Extract relations and their arguments from unstructured text in unsupervised manner

"AT&T, which is based in Dallas, is a telecommunication company."

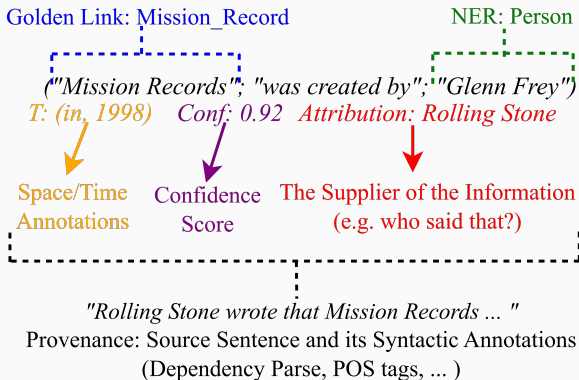


("AT&T"; "is based in"; "Dallas")
("AT&T"; "is"; "telecommunication company")

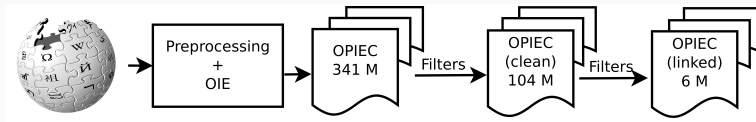
- Big text corpora can produce millions of OIE triples
 - valuable resources for many downstream tasks
 - e.g. automated KB construction, open question answering, event schema induction, ...

OPIEC: An Open Information Extraction Corpus

- The **largest OIE corpus** to date (341M triples)
- **Rich with meta-data**: many syntactic/semantic annotations
- Ran an OIE system on the **entire English Wikipedia**
 - the original **golden links** from a Wikipedia article are kept



Subcorpora: OPIEC-Clean and OPIEC-Linked



- **OPIEC-Clean (104M triples):** Triples whose arguments are **self-contained** and **refer to concepts**
- **OPIEC-Linked (6M triples):** Triples with **linked arguments**

("Michael Jordan"; "grew up in"; "Wilmington")



Analysis: OPIEC and Knowledge Bases

- **Goal:** compare OPIEC triples to KB triples
- OPIEC-Linked triple has a **KB hit** when potentially present in KB (optimistic measure)
 - 70.3% of the linked triples do not have a KB hit
 - OIE facts often differ in the **level of specificity** compared to KB facts

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

Table 1: The most frequent open relations aligned to DBpedia relations

Take-aways

- OPIEC: the **largest OIE corpus** to date
- Aims to spur research in **AKBC, open Q&A, ...**
- **Rich with meta-data:** many syntactic/semantic annotations
- Multiple **sub-corpora** from noisy to clean
- Analyzed and compared with **Wikipedia-based KBs**

Thank you for your attention!