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 an 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

Golden Link: Mission_Record

("Mission Records"; "was created by"; "Glenn Frey")

T: (in, 1998)  Conf: 0.92  Attribution: Rolling Stone

Space/Time Annotations  Confidence Score  The Supplier of the Information (e.g. who said that?)

"Rolling Stone wrote that Mission Records ... "

Provenance: Source Sentence and its Syntactic Annotations
(Dependency Parse, POS tags, ... )
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")
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

<table>
<thead>
<tr>
<th>associatedMusicalArtist</th>
<th>spouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>“be”</td>
<td>(5,521)</td>
</tr>
<tr>
<td>“have”</td>
<td>(3,248)</td>
</tr>
<tr>
<td>“be guitarist of”</td>
<td>(619)</td>
</tr>
<tr>
<td>“be drummer of”</td>
<td>(433)</td>
</tr>
<tr>
<td>“be feature”</td>
<td>(377)</td>
</tr>
<tr>
<td>“be frontman of”</td>
<td>(367)</td>
</tr>
<tr>
<td>“be wife of”</td>
<td>(1,580)</td>
</tr>
<tr>
<td>“be”</td>
<td>(980)</td>
</tr>
<tr>
<td>“marry”</td>
<td>(551)</td>
</tr>
<tr>
<td>“be widow of”</td>
<td>(392)</td>
</tr>
<tr>
<td>“be marry to”</td>
<td>(246)</td>
</tr>
<tr>
<td>“have”</td>
<td>(244)</td>
</tr>
</tbody>
</table>

**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!