OPIEC: An Open Information Extraction Corpus
Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, Rainer Gemulla
University of Mannheim, Mannheim, Germany

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

OPIEC: Open Information Extraction Corpus

OPIEC-Raw (341M Triples)
Constructed using the state-of-the-art OIE system MinIE-SpaTe on Wikipedia

Example Triple:
Golden Link: Mission_Record
"Mission Records", "was created by": "Glenn Frey"
T: (in, 1998)  Conf: 0.92  Attribution: Rolling Stone
The Supplier of the Information (e.g. who said that?)

Space/Time Annotations  Confidence Score
"Rolling Stone wrote that Mission Records ... "
Provenance: Source Sentence and its Syntactic Annotations

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

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

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