Revisiting the Task of Scoring Open IE Relations

William Léchelle, Philippe Langlais
RALI – Université de Montréal, Québec, Canada
rali.iro.umontreal.ca

Protocol

Language Model

- We score unseen triples using a Language Model (trained on the KB, with off-the-shelf KenLM).
- LM-basic: the score is the probability of the span: \( \text{score}(\{a_1, r, a_2\}) = \log p(a_1, r, a_2) \)
- LM-junctions: \( \text{score}(\{a_1, r, a_2\}) = \log p(a_1, r, a_2) - \log p(a_1) - \log p(r) - \log p(a_2) \)
- LM-SVM: a classifier based on 20 LM-based features

Results

<table>
<thead>
<tr>
<th></th>
<th>Reverb-15M</th>
<th>Reverb-500M†</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArgSim</td>
<td>53.0</td>
<td>52.6</td>
</tr>
<tr>
<td>AM-count</td>
<td>61.6</td>
<td>52.3</td>
</tr>
<tr>
<td>AM-cos</td>
<td>64.2</td>
<td>70.6</td>
</tr>
<tr>
<td>AM-system</td>
<td>65.4</td>
<td>74.2</td>
</tr>
<tr>
<td>LM-basic</td>
<td>73.6</td>
<td>76.3</td>
</tr>
</tbody>
</table>

† Reverb-500M is a bigger and noisier superset of Reverb-15M. Results from (Angeli and Manning 2013) on Reverb-500M are presented for comparison.

Context

- Knowledge Base Completion for Open Information Extraction-induced KBs
- (Angeli and Manning, 2013) propose a triple classification protocol
- We propose a Language Model baseline and explore the task setting

Human performance on the task

- Two human experts classified 2x100 triples.
- Both got 80% accuracy, with 80% agreement.
- Examples:
  - (Kara; is vice-president of; buying)
  - (Zoroastrianism; is in even; worse shape)
  - (Cooking school; have changed a bit in; the Los Angeles area)

Classification of a manually labeled test set

- 430 genuine extractions manually labeled as good, bad, or ambiguous
- Classification between the good and bad classes alone yields disappointing results.

Acknowledgements

- Fabrizio Gotti, for invaluable discussion
- Nuance Foundation, for supporting this work

Reverb-15M

We used the Reverb-15M KB. It was obtained by running the Reverb Open IE system on Clueweb09 and filtering the results, resulting in a high quality dataset. It is presented and available at http://reverb.cs.washington.edu/.

Negative sampling knob

- “Negative knob”: how often to pick \((a_1', r, a_2)\) rather than \((a_1, r, a_2')\) or \((a_1', r, a_2')\) for negative examples...
- … has a major impact on the results.

Results from (Angeli and Manning 2013) on Reverb-500M are presented for comparison.

An swers: T, T, F

Performance depending on negative knob

<table>
<thead>
<tr>
<th></th>
<th>LM-SVM</th>
<th>LM-junctions</th>
<th>LM-basic</th>
<th>AM-count</th>
<th>AM-cos</th>
<th>AM-system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reverb-15M</td>
<td>66.5</td>
<td>73.6</td>
<td>71.4</td>
<td>61.6</td>
<td>53.0</td>
<td>52.6</td>
</tr>
<tr>
<td>Reverb-500M†</td>
<td>60.6</td>
<td>66.5</td>
<td>67.4</td>
<td>52.6</td>
<td>49.5</td>
<td>51.5</td>
</tr>
</tbody>
</table>

RV15M-430

- ArgSim: 62.5%
- AM-count: 49.5%
- AM-cos: 51.1%
- LM-basic: 53.3%
- LM-junctions: 51.5%

An swers: T, T, F

An swers: T, T, F