Coverage of Information Extraction from Sentences and Paragraphs

Simon Razniewski¹, Nitisha Jain², Paramita Mirza¹, Gerhard Weikum¹

¹Max Planck Institute for Informatics
²Hasso Plattner Institute
The classical IE paradigm

Obama has two children, Malia and Sasha.

Information extraction tool

<Obama, child, Malia>
<Obama, child, Sasha>

But are these all?
Coverage differences

• Obama has two children, Malia and Sasha.
• Jolie brought her children Shiloh and Knox to school.

• New York consists of the districts Manhattan, Bronx, Queens, Brooklyn, and Staten Island.
• Important districts of Hong Kong are Wan Chai, Kowloon City, and Yau Tsim Mong.
Relevance (1/3): IE resource efficiency

**Districts (NY) = Manhattan, Bronx, Queens, Brooklyn, Staten Island**
Coverage = **High** → Stop further extraction

**Districts (Hong Kong) = Wan Chai, Kowloon City, Yau Tsim Mong**
Coverage = **Low** → Explore more resources
Relevance (2/3): Adjust IE thresholds

- District(HK, Wan Chai) - confidence 0.93
- District(HK, Kowloon City) - confidence 0.86
- District(HK, Yau Tsim) - confidence 0.74
- District(HK, Macao) - confidence 0.67
- ...

HK consists of the districts Wan Chai, ..., ..., ..., ..., and ....
Coverage 0.98
Relevance (3/3): QA negation and completeness

• Which US presidents were married only once?
• Which countries participated in no UN mission?
• For which cities do we know all districts?

Without coverage awareness, QA systems cannot answer these
Focus of our research [SIGMOD’15, WSDM’17, ACL’17, ISWC’18, ...]
Coverage estimation how?

Grice’s maxims of cooperative communication
[Logic and conversation, 1975]

Maxim of quantity:
- Make your contribution as informative as required

Maxim of relevance:
- Be relevant

Obama has two children, Malia and Sasha.

Jolie brought her children Shiloh and Knox to school.

Can we automatically determine where full coverage is relevant?
Formal problem: Full coverage prediction

Subject $s$
Predicate $p$
Real-world set of $p$-objects for $s$: $RW\{o \mid sp\}$
Ground truth extraction (perfect IE): $GTE\{o \mid sp, t\}$

Given a text segment $t$:

$GTE\{o \mid sp, t\} = RW\{o \mid sp\}$?

If yes $t$ has full coverage of $o$ for $s$, $p$
Setup

• \( p \in \{ \text{child, spouse, bandmember, educatedAt, employer} \} \)
• \( s \in \) popular entities from Wikidata
• \( t: \) Wikipedia sentences/paragraphs

• \( \text{GET}_{o | sp, t} \): OpenIE + predicate dictionary + surface name matching
• \( \text{RW}_{o | sp} \): Distant supervision
  • Wikidata objects – assumed to be complete for popular \( s \)

• **Classifier/features:**
  • SVM on text n-grams, LSTM on word embeddings

• **Baselines:**
  • Random
  • Longest text segments are complete
  • Text segments containing most proper names are complete
## Results

<table>
<thead>
<tr>
<th>Text unit</th>
<th>Model</th>
<th>child</th>
<th>bandMember</th>
<th>educatedAt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence</td>
<td>Random</td>
<td>.06</td>
<td>.06</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>.05</td>
<td>.13</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>#pnames</td>
<td>.05</td>
<td>.17</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>.45</td>
<td>.60</td>
<td>.64</td>
</tr>
<tr>
<td>Paragraph</td>
<td>Random</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Length</td>
<td>.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#pnames</td>
<td>.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LSTM</td>
<td>.55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Baselines capture some signal
2. LSTM on text contains much stronger signals
3. Task is not easy

F1-score on predicting text units w/ full coverage.
## Paragraph features

<table>
<thead>
<tr>
<th>child</th>
<th>spouse</th>
<th>bandMember</th>
<th>educatedAt</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;num&gt; grandsons</td>
<td>married twice</td>
<td>consists of</td>
<td>briefly attended</td>
</tr>
<tr>
<td>&lt;pname&gt; sons</td>
<td>second marriage</td>
<td>vocals &lt;pname&gt;</td>
<td>left graduating</td>
</tr>
<tr>
<td>daugthers:</td>
<td>later married</td>
<td>lineup &lt;pname&gt;</td>
<td>&lt;pname&gt; left</td>
</tr>
<tr>
<td>&lt;pname&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example predictions

<table>
<thead>
<tr>
<th>Sentence</th>
<th>LSTM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>He was the father of actor Pierre Renoir (1885-1952), filmmaker Jean Renoir (1894-1979) and ceramic artist Claude Renoir (1901-1969).</td>
<td>0.54</td>
</tr>
<tr>
<td>His daughter Julie Gavras and his son Romain Gavras are also filmmakers.</td>
<td>0.46</td>
</tr>
<tr>
<td>Genghis Khan was aware of the friction between his sons (particularly between Chagatai and Jochi) and worried of possible conflict between them if he died.</td>
<td>0.42</td>
</tr>
<tr>
<td>“From this moment I am no longer the king; the king is Victor my son.”</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Take-home

• IE so far only focused on confidence (precision)
• Coverage (recall) has importance for resource efficiency, thresholding, QA w/ negation
• Linguistic theories give handles towards coverage estimation

• Experiments:
  • Coverage estimation is feasible
  • N-grams provide informative features