Creating a Gold Benchmark for Open IE

Gabi Stanovsky and Ido Dagan
Bar-Ilan University
In this talk

• **Problem**: No large benchmark for Open IE evaluation!

• **Approach**
  • Identify common extraction principles
  • Extract a large Open IE corpus from QA-SRL
  • Automatic system comparison

• **Contributions**
  • Novel methodology for compiling Open IE test sets
  • New corpus readily available for future evaluations
Problem:
Evaluation of Open IE
Open Information Extraction

• Extracts SVO tuples from texts
  • Barack Obama, the U.S president, was born in Hawaii → (Barack Obama, born in, Hawaii)
  • Obama and Bush were born in America → (Obama, born in, America), (Bush, born in, America)

• Useful for populating large databases
  • A scalable open variant of Information Extraction
Open IE: Many parsers developed

- TextRunner (Banko et al., NAACL 2007)
- WOE (Wu and Weld, ACL 2010)
- ReVerb (Fader et al., 2011)
- OLLIE (Mausam et al., EMNLP 2012)
- KrakeN (Akbik and Luser, ACL 2012)
- ClausIE (Del Corro and Gemulla, WWW 2013)
- Stanford Open Information Extraction (Angeli et al., ACL 2015)
- DEFIE (Bovi et al., TACL 2015)
- Open-IE 4 (Mausam et al., ongoing work)
- PropS-DE (Falke et al., EMNLP 2016)
- NestIE (Bhutani et al., EMNLP 2016)
Problem: Open IE evaluation

• Open IE task formulation has been lacking formal rigor
  • No common guidelines → **No large corpus for evaluation**

• Post-hoc evaluation:
  • Annotators judge *a small sample* of their output

→ **Precision oriented** metrics

→ Figures are **not comparable**

→ Experiments are **hard to reproduce**
Previous evaluations

<table>
<thead>
<tr>
<th>System</th>
<th>#Sentences</th>
<th>Genre</th>
<th>Metric</th>
<th>#Annot.</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextRunner</td>
<td>400</td>
<td>Web</td>
<td>% Correct</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>WOE</td>
<td>300</td>
<td>Web, Wiki, News</td>
<td>Precision / Recall</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>ReVerb</td>
<td>500</td>
<td>Web</td>
<td>Precision / AUC</td>
<td>2</td>
<td>86%, .68 k</td>
</tr>
<tr>
<td>KrakeN</td>
<td>500</td>
<td>Web</td>
<td>% Correct</td>
<td>2</td>
<td>87%</td>
</tr>
<tr>
<td>Ollie</td>
<td>300</td>
<td>News, Wiki, Biology</td>
<td>Precision/Yield AUC</td>
<td>2</td>
<td>96%</td>
</tr>
<tr>
<td>ClauseIE</td>
<td>300</td>
<td>Web, Wiki, News</td>
<td>Precision/Yield</td>
<td>2</td>
<td>57% / 68% / 63%</td>
</tr>
</tbody>
</table>

→ Hard to draw general conclusions!
Solution:
Common Extraction Principles
Large Open IE Benchmark
Automatic Evaluation
Common principles

1. **Open lexicon**

2. **Soundness**
   
   “Cruz refused to endorse Trump”
   
   ReVerb: (Cruz; endorse; Trump)
   
   OLLIE: (Cruz; refused to endorse; Trump)

3. **Minimal argument span**
   
   “Hillary promised better education, social plans and healthcare coverage”
   
   ClausIE: (Hillary, promised, better education), (Hillary, promised, better social plans),
   (Hillary, promised, better healthcare coverage)
Solution:

Common Extraction Principles

Large Open IE Benchmark

QA-SRL → Open IE

Automatic Evaluation
## Open IE vs. traditional SRL

<table>
<thead>
<tr>
<th></th>
<th>Open IE</th>
<th>Traditional SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open lexicon</td>
<td>V</td>
<td>X</td>
</tr>
<tr>
<td>Soundness</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Reduced arguments</td>
<td>V</td>
<td>X</td>
</tr>
</tbody>
</table>
Recently, He et al. (2015) annotated SRL by asking and answering argument role questions. Obama, the U.S president, was born in Hawaii.

- Who was born somewhere?  
  Obama

- Where was someone born?  
  Hawaii
<table>
<thead>
<tr>
<th></th>
<th>Open IE</th>
<th>Traditional SRL</th>
<th>QA-SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open lexicon</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>Consistency</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Reduced arguments</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
</tbody>
</table>

QA-SRL isn’t limited to a lexicon

QA-SRL format solicits reduced arguments (Stanovsky et al., ACL 2016)
Converting QA-SRL to Open IE

• Intuition: generate all independent extractions

• Example:
  • “Barack Obama, the newly elected president, flew to Moscow on Tuesday”
  • QA-SRL:
    • Who **flew** somewhere? Barack Obama / the newly elected president
    • Where did someone **fly**? to Moscow
    • When did someone **fly**? on Tuesday

  → OIE: (Barack Obama, **flew**, to Moscow, on Tuesday)
  (the newly elected president, **flew**, to Moscow, on Tuesday)

→ Cartesian product over all answer combinations
  • Special cases for nested predicates, modals and auxiliaries
Resulting Corpus

<table>
<thead>
<tr>
<th>Corpus</th>
<th>WSJ</th>
<th>WIKI</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Sentences</td>
<td>1241</td>
<td>1959</td>
<td>3200</td>
</tr>
<tr>
<td>#Predicates</td>
<td>2020</td>
<td>5690</td>
<td>7710</td>
</tr>
<tr>
<td>#Questions</td>
<td>8112</td>
<td>10798</td>
<td>18910</td>
</tr>
<tr>
<td>#Extracts</td>
<td>4481</td>
<td>5878</td>
<td>10359</td>
</tr>
</tbody>
</table>

- Validated against an expert annotation of 100 sentences (95% F1)
- 13 times bigger than largest previous OIE corpus (ReVerb)
Solution:
Common Extraction Principles
Large Open IE Benchmark
Automatic Evaluation
Evaluation

• We evaluate 6 publicly available systems

  1. ClausIE
  2. Open-IE 4
  3. OLLIE
  4. PropS IE
  5. ReVerb
  6. Stanford Open IE

• Soft matching function to accommodate system flavors
Evaluation

Low recall: Missed long-range dep, pronoun resolution

Stanford’s performance:
Probability of 1 to most extractions
“Duplicates” hurt precision
Caveat

• OIE parsers didn’t tune for our corpus
  ➔ Evaluation may not reflect optimal performance

• More importantly – using our corpus for future system development
Conclusion

• **New benchmark published**
  • [https://github.com/gabrielStanovsky/oie-benchmark](https://github.com/gabrielStanovsky/oie-benchmark)
  • 13 times larger than previous benchmarks

• First automatic and objective OIE evaluation

• Novel method for creating OIE test sets for new domains

Thanks for listening!