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

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

→ **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

	Open IE	Traditional SRL
Open lexicon	V	X
Soundness	V	V
Reduced arguments	V	X

QA-SRL

- 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

Open IE vs. SRL vs. QA-SRL

QA-SRL isn't limited to a lexicon

	Open IE	Traditional SRL	QA-SRL
Open lexicon	V	X	V
Consistency	V	V	V
Reduced arguments	V	X	V

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

Corpus	WSJ	WIKI	All
#Sentences	1241	1959	3200
#Predicates	2020	5690	7710
#Questions	8112	10798	18910
#Extractions	**4481**	**5878**	**10359**

- 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

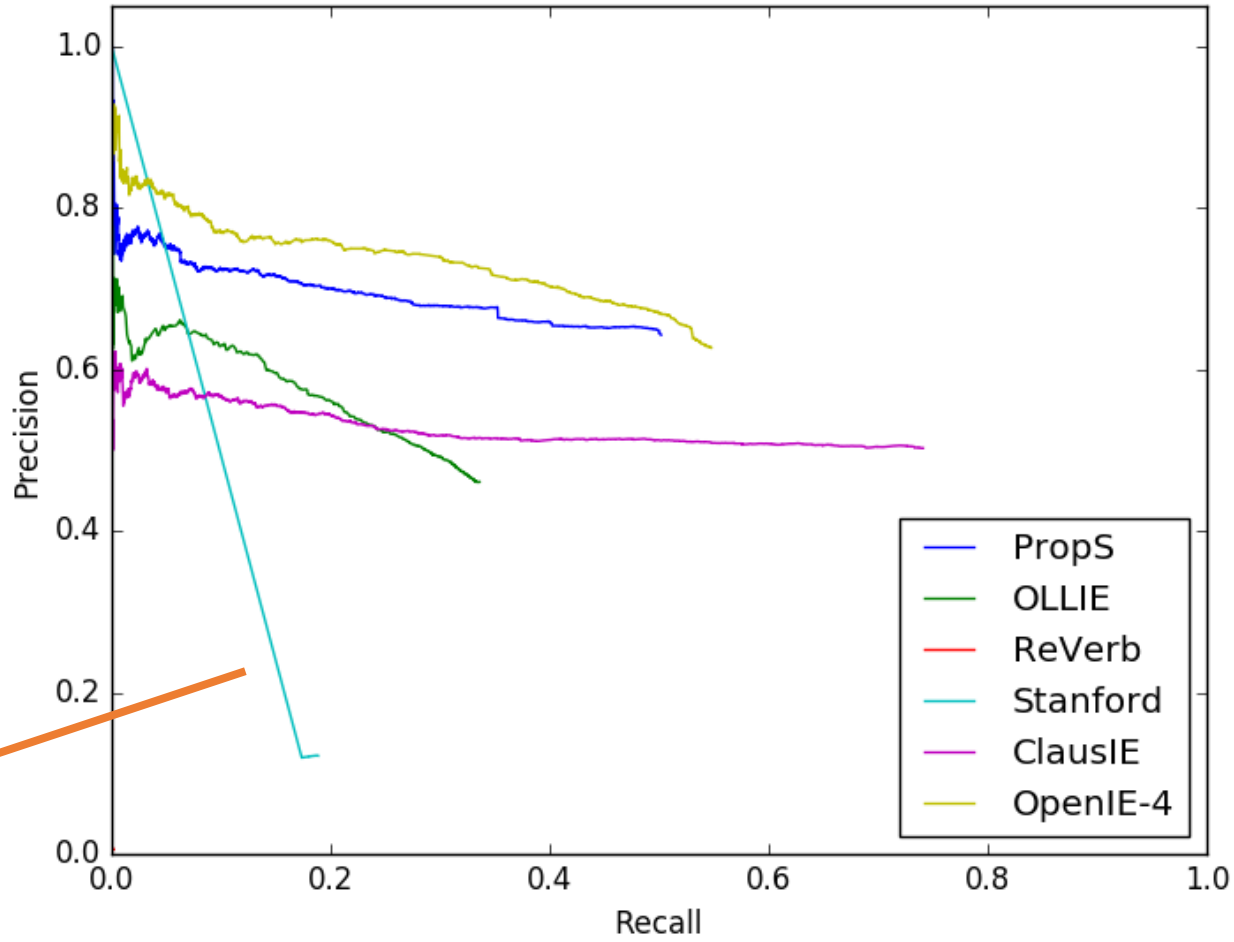
- We evaluate 6 publicly available systems

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

- Soft matching function to accomodate 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
  - 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!