Open Information Extraction from the Web

Oren Etzioni
KnowItAll Project (2003...)

Rob Bart
Janara Christensen
Tony Fader
Tom Lin
Alan Ritter
Michael Schmitz
Dr. Niranjan Balasubramanian
Dr. Stephen Soderland
Prof. Mausam
Prof. Dan Weld

**PhD alumni:** Michele Banko, Prof. Michael Cafarella, Prof. Doug Downey, Ana-Maria Popescu, Stefan Schoenmackers, and Prof. Alex Yates

**Funding:** DARPA, IARPA, NSF, ONR, Google.
Outline

I. A “scruffy” view of Machine Reading
II. Open IE (overview, progress, new demo)
III. Critique of Open IE
IV. Future work: Open, Open IE
I. Machine Reading (Etzioni, AAAI ‘06)

• “MR is an exploratory, open-ended, serendipitous process”

• “In contrast with many NLP tasks, MR is inherently unsupervised”

• “Very large scale”

• “Forming Generalizations based on extracted assertions”
I. Machine Reading (Etzioni, AAAI ’06)

• “MR is an exploratory, open-ended, serendipitous process.”

• “In contrast with many NLP tasks, MR is inherently unsupervised.”

• “Very large scale”

• “Forming Generalizations based on extracted assertions”

Ontology Free!
Lessons from DB/KR Research

• Declarative KR is expensive & difficult
• Formal semantics is at odds with
  – Broad scope
  – Distributed authorship
• KBs are brittle: “can only be used for tasks whose knowledge needs have been anticipated in advance” (Halevy IJCAI ‘03)
Lessons from DB/KR Research

• Declarative KR is expensive & difficult
• Formal semantics is at odds with
  – Broad scope
  – Distributed authorship
• KBs are brittle:
  “can only be used for tasks whose needs have been anticipated”

(Halevy IJCAI '03)

A fortiori, for KBs extracted from text!
Machine Reading at Web Scale

• A “universal ontology” is impossible
• Global consistency is like world peace
• Micro ontologies--scale? Interconnections?

• Ontological “glass ceiling”
  – Limited vocabulary
  – Pre-determined predicates
  – Swamped by reading at scale!
II. Open vs. Traditional IE

<table>
<thead>
<tr>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong></td>
<td>Corpus</td>
</tr>
<tr>
<td><strong>Relations:</strong></td>
<td>Discovered automatically</td>
</tr>
<tr>
<td><strong>Extractor:</strong></td>
<td>Relation-independent</td>
</tr>
<tr>
<td>Corpus + (O(R)) hand-labeled data</td>
<td></td>
</tr>
<tr>
<td>Specified in advance</td>
<td></td>
</tr>
<tr>
<td>Relation-specific</td>
<td></td>
</tr>
</tbody>
</table>

How is Open IE Possible?
Semantic Tractability Hypothesis

\[ \exists \textit{easy-to-understand} \text{ subset of English} \]

- Characterized relations/arguments syntactically
  (Banko, ACL ’08; Fader, EMNLP ’11; Etzioni, IJCAI ’11)

- Characterization is compact, domain independent

- Covers 85% of binary, verb-based relations
<table>
<thead>
<tr>
<th>Action</th>
<th>By/For</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>invented</td>
<td>acquired by</td>
<td>has a PhD in</td>
</tr>
<tr>
<td>denied</td>
<td>voted for</td>
<td>inhibits tumor growth in</td>
</tr>
<tr>
<td>inherited</td>
<td>born in</td>
<td>mastered the art of</td>
</tr>
<tr>
<td>downloaded</td>
<td>aspired to</td>
<td>is the patron saint of</td>
</tr>
<tr>
<td>expelled</td>
<td>Arrived from</td>
<td>wrote the book on</td>
</tr>
</tbody>
</table>
## Number of Relations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DARPA MR Domains</td>
<td>&lt;50</td>
</tr>
<tr>
<td>NYU, Yago</td>
<td>&lt;100</td>
</tr>
<tr>
<td>NELL</td>
<td>~500</td>
</tr>
<tr>
<td>DBpedia 3.2</td>
<td>940</td>
</tr>
<tr>
<td>PropBank</td>
<td>3,600</td>
</tr>
<tr>
<td>VerbNet</td>
<td>5,000</td>
</tr>
<tr>
<td>WikiPedia InfoBoxes, f &gt; 10</td>
<td>~5,000</td>
</tr>
<tr>
<td>TextRunner (phrases)</td>
<td>100,000+</td>
</tr>
<tr>
<td>ReVerb (phrases)</td>
<td>1,000,000+</td>
</tr>
</tbody>
</table>

Etzioni, University of Washington
TextRunner (2007)

First Web-scale Open IE system
Distant supervision + CRF models of relations

(Arg1, Relation phrase, Arg2)

1,000,000,000 distinct extractions

Etzioni, University of Washington
Relation Extraction from Web

![Graph showing precision and recall for different models: ReVERB, WOEparse, WOEPOS, TextRUNNER. The graph displays the performance comparison with varying recall values.](image-url)
Open IE (2012)

- Open source ReVerb extractor
- Synonym detection
- Parser-based Ollie extractor *(Mausam EMNLP ‘12)*
  - Verbs ➔ Nouns and more
  - Analyze context (beliefs, counterfactuals)
- Sophistication of IE is a major focus

But what about entities, types, ontologies?
• Open source
• Synonym detection
• Parser-based
  – Verbs ➔ Nouns and more
  – Analyze context (beliefs, counterfactuals)
• Sophistication of IE is a major focus

But what about entities, types, ontologies?

After beating the Heat, the Celtics are now the “top dog” in the NBA. (the Celtics, beat, the Heat)
If he wins 5 key states, Romney will be president.

(counterfactual: “if he wins 5 key states”)

But what about entities, types, ontologies?
Towards “Ontologized” Open IE

• Link arguments to Freebase (Lin, AKBC ‘12)
  – When possible!
• Associate types with Args
• No Noun Phrase Left Behind
  (Lin, EMNLP ‘12)
## System Architecture

<table>
<thead>
<tr>
<th>Input</th>
<th>Processing</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web corpus</td>
<td>Extractor</td>
<td>(XYZ Corp.; acquired; Go Inc.) (oranges; contain; Vitamin C) (Einstein; was born in; Ulm) (XYZ; buyout of; Go Inc.) (Albert Einstein; born in; Ulm) (Einstein Bros.; sell; bagels)</td>
</tr>
<tr>
<td>Raw tuples</td>
<td>Assessor</td>
<td>XYZ Corp. = XYZ Albert Einstein = Einstein != Einstein Bros.</td>
</tr>
</tbody>
</table>

**Relation-independent extraction**

**Synonyms, Confidence**

**Index in Lucene; Link entities**

---

Etzioni, University of Washington
III. Critique of Open IE

• Lack of formal ontology/vocabulary
• Inconsistent extractions
• Can it support reasoning?
• What’s the point of Open IE?
Perspectives on Open IE

A. “Search Needs a Shakeup” (Etzioni, Nature ’11)
B. Textual Resources
C. Reasoning over Extractions
A. New Paradigm for Search

“Moving Up the Information Food Chain”
(Etzioni, AAAI ‘96)

- Retrieval ➔ Extraction
- Snippets, docs ➔ Entities, Relations
- Keyword queries ➔ Questions
- List of docs ➔ Answers

Essential for smartphones!
(Siri meets Watson)
Case Study over Yelp Reviews

1. Map review corpus to (attribute, value)
   (sushi = fresh) (parking = free)
2. Natural-language queries
   “Where’s the best sushi in Seattle?”
3. Sort results via sentiment analysis
   exquisite > very good > so, so
RevMiner: Extractive Interface to 400K Yelp Reviews (Huang, UIST ’12)
RevMiner: Extractive Interface to 400K Yelp Reviews (Huang, UIST ’12)
B. Public Textual Resources (Leveraging Open IE)

- **94M** Rel-grams: n-grams, but over relations in text (Balasubramanian. AKBC’12)
- **600K** Relation phrases (Fader, EMNLP ‘11)
- Relation Meta-data:
  - **50K** Domain/range for relations (Ritter, ACL ‘10)
  - **10K** Functional relations (Lin, EMNLP ‘10)
- **30K** learned Horn clauses (Schoenmackers, EMNLP ‘10)
- **CLEAN** (Berant, ACL ‘12)
  - **10M** entailment rules (coming soon)
  - Precision double that of DIRT

See openie.cs.washington.edu
B. Public Textual Resources

(Leveraging Open IE)

- **94M** Rel-grams: n-grams, but over relations in text (Balasubramanian)
- **600K** Relation phrases (Fader, EMNLP ‘11)
- **Relation Meta-data:**
  - **50K** Domain/range for relations (Ritter, ACL ‘10)
  - **10K** Functional relations (Lin, EMNLP ‘10)
- **30K** learned Horn clauses (Schoenmackkers, EMNLP ‘10)
- **CLEAN** (Berant, ACL ‘12)
  - **10M** entailment rules (coming soon)
  - Precision double that of DIRT

See openie.cs.washington.edu

Etzioni, University of Washington
C. Reasoning over Extractions

1,000,000,000 Extractions

- Identify synonyms
  (Yates & Etzioni JAIR ‘09)

- Linear-time 1st order Horn-clause inference
  (Schoenmackers EMNLP ‘08)

- Learn argument types
  Via generative model
  (Ritter ACL ‘10)

- Transitive Inference
  (Berant ACL ‘11)
Unsupervised, probabilistic model for identifying synonyms

• $P(\text{Bill Clinton} = \text{President Clinton})$
  – Count shared (relation, arg2)

• $P(\text{acquired} = \text{bought})$
  – Relations: count shared (arg1, arg2)

• Functions, mutual recursion

• Next step: unify with

Etzioni, University of Washington
Scalable Textual Inference

Desiderata for inference:
• In text $\rightarrow$ probabilistic inference
• On the Web $\rightarrow$ linear in $|\text{Corpus}|$

Argument distributions of textual relations:
• Inference provably linear
• **Empirically linear!**
Inference Scalability for Holmes
Extractions ➔ Domain/range

• Much previous work (Resnick, Pantel, etc.)
• Utilize generative topic models

Extractions of R ➔ Document
Domain/range of R ➔ topics
born_in(Sergey Brin, Moscow)
headquartered_in(Microsoft, Redmond)

born_in(Bill Gates, Seattle)

born_in(Einstein, March)
founded_in(Google, 1998)

headquartered_in(Google, Mountain View)
born_in(Sergey Brin, 1973)
founded_in(Microsoft, Albuquerque)
born_in(Einstein, Ulm)
founded_in(Microsoft, 1973)
Generative Story
[LinkLDA, Erosheva et al. 2004]

\[ X \text{ born\_in } Y \]
\[ P(\text{Topic1}|\text{born\_in})=0.5 \]
\[ P(\text{Topic2}|\text{born\_in})=0.3 \]
...

**Person** born\_in **Location**

Sergey Brin born\_in Moscow

For each relation, randomly pick a distribution over types

For each extraction, pick type for a1, a2

Then pick arguments based on types

Two separate sets of type distributions
Examples of Learned Domain/range

- **elect**(Country, Person)
- **predict**(Expert, Event)
- **download**(People, Software)
- **invest**(People, Assets)
- **Was-born-in**(Person, Location OR Date)
Summary: Trajectory of Open IE

- 2003: KnowItAll project
- 2007: TextRunner: 1,000,000,000 “Ontology free” extractions
- 2008-9: Inference over extractions
- 2010-11: Open source extractor
  - Public textual Resources
- 2012: Freebase types
  - IE-based search
  - Deeper analysis of sentences

Openie.cs.washington.edu

Etzioni, University of Washington
IV. Future: Open Open IE

• **Open input:** ingest tuples from any source (Tuple, Source, Confidence)

• **Linked Open Output:**
  – Extractions ➔ Linked-open Data (LOD) cloud
  – Relation normalization
  – Use LOD best practices

• **Specialized reasoners**
Conclusions

1. Ontology is not necessary for reasoning
2. Open IE is “gracefully” ontologized
3. Open IE is boosting text analysis
4. LOD has distribution & scale (but not text) = opportunity

Thank you