Open Information Extraction Systems
and Downstream Applications

Mausam
Computer Science and Engineering
Indian Institute of Technology
New Delhi, India

Joint work with
Oren Etzioni, Stephen Soderland, Michael Schmitz,
Ido Dagan, Ganesh Ramakrishnan, Sunita Sarawagi, Parag Singla,
Niranjan Balasubramanian, Robert Bart, Janara Christensen,
Danish Contractor, Anthony Fader, Aman Madaan, Ashish Mittal,
Harinder Pal, Abhishek Yadav, Gabriel Stanovsky
“The Internet is the world’s largest library. It’s just that all the books are on the floor.”

- John Allen Paulos

~20 Trillion URLs (Google)
Closed KBs

Populating information wrt a given ontology from natural language text

"Apple’s founder Steve jobs died of cancer following a..."

Closed IE

rel:founder_of(Apple, Steve Jobs)

rel:founder_of
(Apple, Steve Jobs)
(Google, Larry Page)
(Facebook, Mark Zuckerberg)

rel:acquisition
(Google, DeepMind)
(IBM, uStream)
(Microsoft, Powerset)

...
Lessons from DB/KR Research

• Large-scale Ontologies
  – expensive to create
  – very hard to maintain
  – conflict with distributed authorship

• KBs are brittle: “can only be used for tasks whose knowledge needs have been anticipated in advance” (Halevy IJCAI ‘03)
Open KBs

Broad-coverage KBs with light-weight structure
(no annotation per relation)

“When Saddam Hussain invaded Kuwait in 1990, the international..”

(Saddam Hussain, invaded, Kuwait)

(Google, acquired, Youtube)
(Oranges, contain, Vitamin C)
(Edison, invented, phonograph)
...

antibiotics (381)
Chlorine (113)
Ozone (61)
Heat (60)
Honey (55)
Benzoyl peroxide (45)

The heat kills the bacteria.
The heat kills bacteria.
Only heat kills bacteria.
Heat kills most bacteria.
Heat can kill the bacteria.
Heat will kill bacteria.
The high heat will kill bacteria.
Heat does kill bacteria.
Demo

Overview

Extraction → Fact → KB

End-user applications

Downstream NLP/AI Tasks

Character icon with megaphone
Open Information Extraction

- **2007: Textrunner (~Open IE 1.0)**
  - CRF and self-training

- **2010: ReVerb (~Open IE 2.0)**
  - POS-based relation pattern

- **2012: OLLIE (~Open IE 3.0)**
  - Dep-parse based extraction; nouns; attribution

- **2014: Open IE 4.0**
  - SRL-based extraction; temporal, spatial...

- **2016 [@IITD]: Open IE 5.0**
  - Compound noun phrases, numbers, lists

Increasing precision, recall, expressiveness
Fundamental Hypothesis

∃ *semantically tractable* subset of English

• Characterized relations & arguments via POS

• Characterization is compact, domain independent

• Covers 85% of binary relations in sample
Identify **Relations** from **Verbs**.

1. Find longest phrase matching a simple syntactic constraint:

\[
\begin{align*}
V & | VP | VW^*P \\
V & = \text{verb particle? adv?} \\
W & = (\text{noun} | \text{adj} | \text{adv} | \text{pron} | \text{det}) \\
P & = (\text{prep} | \text{particle} | \text{inf. marker})
\end{align*}
\]
Sample of ReVerb Relations

- invented
- acquired by
- has a PhD in

- inhibits tumor growth in
- voted in favor of
- won an Oscar for

- has a maximum speed of
- died from complications of
- mastered the art of

- gained fame as
- granted political asylum to
- is the patron saint of

- was the first person to
- identified the cause of
- wrote the book on
## 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>1,500,000+</td>
</tr>
</tbody>
</table>
ReVerb: Error Analysis

• Larry Page, the CEO of Google, talks about multi-screen opportunities offered by Google.

• After winning the Superbowl, the Giants are now the top dogs of the NFL.

• Ahmadinejad was elected as the new President of Iran.

OLLIE: Open Language Learning for Information Extraction
Bootstrapping Approach

Verb-based relations

Other Syntactic rels

Semantic rels
Federer is coached by Paul Annacone.
Bootstrapping Approach

Federer is coached by Paul Annacone.

Now coached by Paul Annacone, Federer has ...

- Reverb's Verb-based relations
- Other Syntactic rels
- Semantic rels
Bootstrapping Approach

Federer is coached by Paul Annacone.

Paul Annacone, the coach of Federer,

Reverb's Verb-based relations

Other Syntactic rels

Semantic rels

Now coached by Paul Annacone, Federer has ...
Federer is coached by Paul Annacone.

Paul Annacone, the coach of Federer,

Reverb’s Verb-based relations

Other Syntactic rels

Semantic rels

Now coached by Paul Annacone, Federer has ...

Federer hired Annacone as his new coach.
Evaluation

[Mausam, Schmitz, Bart, Soderland, Etzioni - EMNLP’12]
Context Analysis

“John refused to visit Vegas.”
(John, visit, Vegas)

“Early astronomers believed that the earth is the center of the universe.”
(earth, is the center of, universe)

“If she wins California, Hillary will be the nominated presidential candidate.”
(Hillary, will be nominated, presidential candidate)
“John refused to visit Vegas.”
(John, refused to visit, Vegas)

“Early astronomers believed that the earth is the center of the universe.”
[(earth, is the center of, universe) Attribution: early astronomers]

“If she wins California, Hillary will be the nominated presidential candidate.”
[(Hillary, will be nominated, presidential candidate) Modifier: if she wins California]
Open Information Extraction

• 2007: Textrunner (~Open IE 1.0)
  – CRF and self-training
• 2010: ReVerb (~Open IE 2.0)
  – POS-based relation pattern
• 2012: OLLIE (~Open IE 3.0)
  – Dep-parse based extraction; nouns; attribution
• 2014: Open IE 4.0
  – SRL-based extraction; temporal, spatial...
• 2016 [@IITD]: Open IE 5.0
  – compound noun phrases, numbers, lists
Compound Noun Extraction

[Pal & Mausam - AKBC’16]

• NIH Director Francis Collins

  (Francis Collins, is the Director of, NIH)

• Challenges
  – New York Banker Association
  – German Chancellor Angela Merkel
  – Prime Minister Modi
  – GM Vice Chairman Bob Lutz
“Venezuela with its inflation rate 96% is suffering from a major...”

(Venezuela, inflation rate, 96 %)
Overview

Useful for
- Summarization
- Comparisons
- Interaction w/ data

Extraction → Fact → KB → End-user applications

Downstream NLP/AI Tasks
Extractions: a great way to summarize
### Extractions: a great way to compare

[Contractor, Mausam, Singla - NAACL’16]

<table>
<thead>
<tr>
<th>Cluster Labels</th>
<th>Granada (Spain)</th>
<th>New York City (U.S.)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>art, arch.</strong></td>
<td>moorish architecture, religious art, fine art, beautiful architecture</td>
<td>contemporary art, modern american art, medieval art, egyptian art</td>
</tr>
<tr>
<td><strong>palace, courtyard</strong></td>
<td>brick-walled courtyard, lovely courtyard area, nasrid royal palace, alhambra palace</td>
<td></td>
</tr>
<tr>
<td>museum, finest</td>
<td>alhambra museum, archaeological museum, world heritage site, splendid arabic shops</td>
<td>fine art museums, guggenheim museum, islamic art collection, metropolitan museum</td>
</tr>
<tr>
<td>gardens, park</td>
<td>partial gardens, palace gardens, pleasant gardens, moorish style gardens</td>
<td>flushing meadows park, central park, renowned gardens, natl. recreational area</td>
</tr>
</tbody>
</table>
Extracts: a great way to compare

[Contractor, Mausam, Singla - NAACL’16]

<table>
<thead>
<tr>
<th>Cluster Labels</th>
<th>Granada (Spain)</th>
<th>New York City (U.S.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>art, arch.</td>
<td>moorish architecture, religious art, fine art, beautiful architecture</td>
<td>contemporary art, modern american art, medieval art, egyptian art</td>
</tr>
<tr>
<td>palace, courtyard</td>
<td>brick-walled courtyard, lovely courtyard area, nasrid royal palace, alhambra palace</td>
<td></td>
</tr>
<tr>
<td>museum, finest</td>
<td>alhambra museum, archaeological museum, world heritage site, splendid arabic shops</td>
<td>fine art museums, guggenheim museum, islamic art collection, metropolitan museum</td>
</tr>
<tr>
<td>gardens, park</td>
<td>partial gardens, palace gardens, pleasant gardens, moorish style gardens</td>
<td>flushing meadows park, central park, renowned gardens, natl. recreational area</td>
</tr>
</tbody>
</table>
Extractions: a great way to compare

[Contractor, Mausam, Singla - NAACL’16]

<table>
<thead>
<tr>
<th>Cluster Labels</th>
<th>Granada (Spain)</th>
<th>New York City (U.S.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>art, arch.</td>
<td>moorish architecture</td>
<td>contemporary art</td>
</tr>
<tr>
<td></td>
<td>religious art</td>
<td>modern american art</td>
</tr>
<tr>
<td></td>
<td>fine art</td>
<td>medieval art</td>
</tr>
<tr>
<td></td>
<td>beautiful architecture</td>
<td>egyptian art</td>
</tr>
<tr>
<td>palace, courtyard</td>
<td>brick-walled courtyard</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lovely courtyard area</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nasrid royal palace</td>
<td></td>
</tr>
<tr>
<td></td>
<td>alhambra palace</td>
<td></td>
</tr>
<tr>
<td>museum, finest</td>
<td>alhambra museum</td>
<td>fine art museums</td>
</tr>
<tr>
<td></td>
<td>archaeological museum</td>
<td>guggenheim museum</td>
</tr>
<tr>
<td></td>
<td>world heritage site</td>
<td>islamic art collection</td>
</tr>
<tr>
<td></td>
<td>splendid arabic shops</td>
<td>metropolitan museum</td>
</tr>
<tr>
<td>gardens, park</td>
<td>partial gardens</td>
<td>flushing meadows park</td>
</tr>
<tr>
<td></td>
<td>palace gardens</td>
<td>central park</td>
</tr>
<tr>
<td></td>
<td>pleasant gardens</td>
<td>renowned gardens</td>
</tr>
<tr>
<td></td>
<td>moorish style gardens</td>
<td>natl. recreational area</td>
</tr>
</tbody>
</table>
Overview

Extraction → Fact → KB

End-user applications
Downstream NLP/AI Tasks
Pipeline for Semantics Tasks

- Tokenization
- Lemmatization
- POS Tagging
- Chunking
- Info Extraction
- SRL
- NER
- Syntactic Parsing

Semantics Tasks
(WSD, QA, Sentiment, Coreference, Events, ...
Pipeline for Semantics Tasks

Sentence Preprocessing

Semantics Tasks
(WSD, QA, Sentiment, Coreference, Events, …)
Abstracted Pipeline

Sentence

Preprocessing

Intermediate Structure (words, parses, SRL frames)

Task-specific model/features

Semantics Tasks
Research Question
[Stanovsky, Dagan, Mausam - ACL’15]

Can Open IE be useful as an alternative intermediate structure?
Lexical Similarity/Analogies

- We experiment by switching **representations**
  - We compute Open IE based embeddings instead of lexical or syntactic context-based embeddings

<table>
<thead>
<tr>
<th>Target</th>
<th>Lexical</th>
<th>Dependency</th>
<th>SRL</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>nsubj_John</td>
<td>A0_John</td>
<td>0_John</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>xcomp_visit</td>
<td>A1_to</td>
<td>1_to</td>
<td></td>
</tr>
<tr>
<td>refused</td>
<td>visit</td>
<td>A1_visit</td>
<td>1_visit</td>
<td></td>
</tr>
<tr>
<td>Vegas</td>
<td></td>
<td>A1_Vegas</td>
<td>2_Vegas</td>
<td></td>
</tr>
</tbody>
</table>
Results

• Lexical similarity

<table>
<thead>
<tr>
<th></th>
<th>Open IE</th>
<th>Lexical</th>
<th>Deps</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>bruni</td>
<td>.757</td>
<td>.735</td>
<td>.618</td>
<td>.491</td>
</tr>
<tr>
<td>luong</td>
<td>.288</td>
<td>.229</td>
<td>.197</td>
<td>.171</td>
</tr>
<tr>
<td>radinsky</td>
<td>.681</td>
<td>.674</td>
<td>.592</td>
<td>.433</td>
</tr>
<tr>
<td>simlex</td>
<td>.39</td>
<td>.365</td>
<td>.447</td>
<td>.306</td>
</tr>
<tr>
<td>ws353-rel</td>
<td>.647</td>
<td>.64</td>
<td>.492</td>
<td>.551</td>
</tr>
<tr>
<td>ws353-sym</td>
<td>.77</td>
<td>.763</td>
<td>.759</td>
<td>.439</td>
</tr>
<tr>
<td>ws353-full</td>
<td>.711</td>
<td>.703</td>
<td>.629</td>
<td>.693</td>
</tr>
</tbody>
</table>

Near-state-of-art for the amount of training data

Functional similarity
Results

• Lexical analogy \( a : a ^ * :: b : b ^ * ? \)

<table>
<thead>
<tr>
<th></th>
<th>Open IE</th>
<th>Lexical</th>
<th>Deps</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google (Add)</td>
<td>0.714</td>
<td>0.651</td>
<td>0.34</td>
<td>0.352</td>
</tr>
<tr>
<td>Google (Mult)</td>
<td>0.729</td>
<td>0.656</td>
<td>0.367</td>
<td>0.362</td>
</tr>
<tr>
<td>MSR (Add)</td>
<td>0.529</td>
<td>0.438</td>
<td>0.4</td>
<td>0.389</td>
</tr>
<tr>
<td>MSR (Mult)</td>
<td>0.55</td>
<td>0.455</td>
<td>0.434</td>
<td>0.406</td>
</tr>
</tbody>
</table>

State of the art

\[
\arg \max_{b^* \in V} \left( \cos(b^*, b) - \cos(b^*, a) + \cos(b^*, a^*) \right)
\]

\[
\arg \max_{b^* \in V} \frac{\cos(b^*, b) \cos(b^*, a^*)}{\cos(b^*, a) + \varepsilon}
\]
Why does Open IE do better?

• **Word Analogy**
  – Captures domain and functional similarity
    *(gentlest: gentler), (loudest: ?)*

• Lexical: higher-pitched  ![X]  [Domain Similar]

• Syntactic: thinnest  ![X]  [Functionally Similar]

• SRL: unbelievable  ![X]  [Functionally Similar?]

• Open-IE: louder  ![✓]
Other NLP Applications

• Atomic relations → Event schemas

• Unsupervised sentence similarity

• Reading comprehension tasks

• Open IE → Closed IE
  – Software (OREO):
    http://homes.cs.washington.edu/~mausam/software.html
Future Work

Extraction → Fact → KB

Inference

End-user applications

Downstream NLP/AI Tasks
Key Future Direction

• Large-scale inference over Open IE

(iron, is a good conductor of, electricity) ↓

(iron nail, conducts, electricity) ↓

(David Beckham, was born in, London) ↓

(David Beckham, was born in, England)
Thanks

Google
IBM Research
DARPA

Bloomberg

USN ONR
DEPARTMENT OF THE NAVY Science & Technology

KISTI
www.kisti.re.kr

IARPA
BE THE FUTURE