We KnowItAll: Lessons from a Quarter Century of Web Extraction Research

Oren Etzioni
Turing Center
University of Washington

http://turing.cs.washington.edu

KnowItAll Team

- Michele Banko
- Michael Cafarella
- Tom Lin
- Mausam
- Alan Ritter
- Stefan Schoenmackers
- Stephen Soderland
- Dan Weld

Alumni: Doug Downey, Ana-Maria Popescu, Alex Yates, and others.
“In the Knowledge Lies the Power”

- Manual “knowledge engineering”
  - Cyc, Freebase, WordNet

- Volunteers
  - Open Mind, DBPedia
  - von Ahn: games ➔ Verbosity

- KnowItAll project: “read” the Web

Outline

1. Turing Center
2. Machine Reading & Open IE
3. TextRunner
4. Textual Inference
5. Conclusions
2. What is Machine Reading?

Unsupervised understanding of text

Information extraction + inference

Information Extraction (IE) via Rules (Hearst '92)

- "...Cities such as Boston, Los Angeles, and Seattle..."

("C such as NP1, NP2, and NP3") => IS-A(each(head(NP)), C), ...

- Detailed information for several countries such as maps, ..."ProperNoun(head(NP))

- "I listen to pretty much all music but prefer country such as Garth Brooks"
IE as Supervised Learning
(E.g., Riloff ‘96, Soderland ‘99)

- Hand-label examples of each relation
- Manual labor linear in |relations|
- Learn lexicalized extractor

Semi-Supervised Learning

- Few hand-labeled examples per relation!
- → Limit on the number of relations
- → Relations are pre-specified
- → Problematic for Machine Reading

- Alternative: self-supervised learning
  - Learner discovers relations on the fly (Sekine ’06)
  - Learner automatically labels examples
  - "Look Ma, No Hands!" (Downey & Etzioni ’08)
**Open IE = Self-supervised IE**  
(Banko, et. al, IJCAI ’07, ACL ’08)

<table>
<thead>
<tr>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong></td>
<td>Corpus + Hand-labeled Data</td>
</tr>
<tr>
<td><strong>Relations:</strong></td>
<td>Specified in Advance</td>
</tr>
<tr>
<td><strong>Complexity:</strong></td>
<td>$O(D \times R)$</td>
</tr>
<tr>
<td><strong>Output:</strong></td>
<td>Lexicalized, relation-specific</td>
</tr>
</tbody>
</table>

---

**How is Open IE Possible?**

- There is a compact set of relationship expressions in English  
  (Banko & Etzioni ACL ’08)

- “Relationship expressions” can be specified without reference to the individual relation!

- Gates founded Microsoft → founded(Gates, Microsoft)


**Relation-Indep. Extraction Rules**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>$E_1 \text{ Verb } E_2 \ X \text{ established } Y$</td>
<td>37.8%</td>
</tr>
<tr>
<td>Noun+Prep</td>
<td>$E_1 \text{ NP Prep } E_2 \ \text{the } X \text{ settlement with } Y$</td>
<td>22.8%</td>
</tr>
<tr>
<td>Verb+Prep</td>
<td>$E_1 \text{ Verb Prep } E_2 \ X \text{ moved to } Y$</td>
<td>16.0%</td>
</tr>
<tr>
<td>Infinitive</td>
<td>$E_1 \text{ to Verb } E_2 \ X \text{ to acquire } Y$</td>
<td>9.4%</td>
</tr>
<tr>
<td>Modifier</td>
<td>$E_1 \text{ Verb } E_2 \text{ NP } \ X \text{ is } Y \text{ winner}$</td>
<td>5.2%</td>
</tr>
<tr>
<td>Coordinate $$E_1 (\text{and</td>
<td>,</td>
<td>-</td>
</tr>
<tr>
<td>Coordinate $$E_1 (\text{and</td>
<td>,}) E_2 \text{ Verb } \ X . Y \text{ merge}$</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Relation-Independent Observations**

- 95% of sample $\Rightarrow$ 8 broad patterns!
- Rules are simplified...
- Applicability conditions complex
  - "$E_1 \text{ Verb } E_2$"
  1. Kentucky Fried Chicken
  2. Microsoft announced Tuesday that...

- Learned via a CRF model
- Effective, but far from perfect!
Lesson 1: to scale IE to the Web corpus, use Open IE

What about extraction errors?
Leveraging the Web’s Redundancy
(Downey, Soderland, & Etzioni, IJCAI ’05)

1) Repetition in distinct sentences
2) Multiple, distinct extraction rules

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Atlanta and other cities”</td>
<td>980</td>
</tr>
<tr>
<td>“Canada and other cities”</td>
<td>286</td>
</tr>
<tr>
<td>“cities such as Atlanta”</td>
<td>5860</td>
</tr>
<tr>
<td>“cities such as Canada”</td>
<td>7</td>
</tr>
</tbody>
</table>

**IJCAI ‘05:** A formal model of these intuitions

---

**IJCAI ’05 Combinatorial Model**

If an extraction $x$ appears $k$ times in a set of $n$ distinct sentences matching a rule, what is the probability that $x \in \text{class } C$?

$$P(x \in C | x \text{ appears } k \text{ times in } n \text{ draws}) = \frac{\sum_{r \in \text{num}(C)} \left( \frac{r}{s} \right)^k \left(1 - \frac{r}{s}\right)^{n-k}}{\sum_{r' \in \text{num}(C \cup E)} \left( \frac{r'}{s} \right)^k \left(1 - \frac{r'}{s}\right)^{n-k}}$$
Lesson 2: utilize the Web’s redundancy

What about complex sentences?

- Paris, which has been proclaimed by many literary figures to be a great source of inspiration, is also a capital city, but not the capital city of an ordinary country, but rather the capital of a great republic---the republic of France!

  Paris is the Capital of France.
But What About...

- Relative clauses
- Disjunction & conjunction
- Anaphora

- Quantification
- Temporal qualification
- Counterfactuals

See Durm & Schubert '08; MacCartney '08

Lesson 3: focus on ‘tractable’ sentences
3. **TextRunner** (Web’s 1st Open IE system)

1. **Self-Supervised Learner**: automatically labels example extractions & trains a CRF-based extractor

2. **Single-Pass Extractor**: makes a single pass over corpus, identifying extractions in ‘tractable’ sentences

3. **Search Engine**: indexes and ranks extractions based on redundancy

4. **Query Processor**: interactive speeds

---

**Application: Information Fusion**

- What kills bacteria?
- What west coast, nano-technology companies are hiring?
- Compare Obama’s “buzz” versus Hillary’s?
- What is a quiet, inexpensive, 4-star hotel in Vancouver?
TextRunner Demo

Extraction run at Google on 500,000,000 high-quality Web pages.
Triples
11.3 million

With Well-Formed Relation
9.3 million

With Well-Formed Entities
7.8 million

Abstract
6.8 million
79.2% correct

Concrete
1.0 million
88.1% correct

Sample of Web Pages

- Concrete facts:
  (Oppenheimer, taught at, Berkeley)

- Abstract facts:
  (fruit contain, vitamins)
Lesson 4: Open IE yields a mass of isolated “nuggets”

Textual inference is the next step towards machine reading

4. Examples of Textual Inference

I. Entity and predicate resolution
II. Composing facts to draw conclusions
I. Entity Resolution

\[ P(\text{String1} = \text{String2})? \]

**Resolver** *(Yates & Etzioni, HLT '07)*: determines synonymy based on relations found by TextRunner (cf. Pantel & Lin '01)

- (X, born in, 1941) (M, born in, 1941)
- (X, citizen of, US) (M, citizen of, US)
- (X, friend of, Joe) (M, friend of, Mary)

\[ P(X = M) \sim \text{shared relations} \]

---

**Relation Synonymy**

- (1, R, 2) (1, Q, 2)
- (2, R 4) (2, Q, 4)
- (4, R, 8) (4, Q, 8)
- Etc. Etc.

\[ P(R = Q) \sim \text{shared argument pairs} \]

- See paper for probabilistic model
- \(O(n \log(n))\) algorithm, \(n = |\text{tuples}|\)
Functional Relations help resolution

1. married-to(Oren, X)
2. married-to(Oren, Y)

If function(married-to)
then P(X=Y) is high

- Caveat: Oren & Oren refer to the same individual

---

Lesson 4: leverage relations implicit in the text

Functional relations are particularly informative
Returning to Our Caveat...

- BornIn(Oren, New York)
- BornIn(Oren, Maryland)
- But these are two different Orens.
- How can you tell?
  - Web page context

Lesson 5: context enhances entity resolution

To date, our extraction process has been myopic
III. Compositional Inference is Key

The *Tour de France* is the world's largest *cycle race*.

The *Tour de France* takes place in *France*.

The illuminated helmet is used during *sporting events such as cycle racing*.

Is-A(*Tour de France*, cycle race)

Information Extraction

In(*Tour de France*, France)

Is-A(*cycle race*, sporting event)

Scalable Textual Inference

(Schoenmackers, Etzioni, Weld EMNLP ‘08)

Desiderata for inference:
- In text ➔ probabilistic inference
- On the Web ➔ scalable
  ➔ linear in |Corpus|

Novel property of textual relations:
- Prevalent
- Provably linear
- *Empirically linear*!
Inference Scalability for Holmes

Related Work

- Weld’s Intelligence in Wikipedia project
- Sekine’s “pre-emptive IE”
- Powerset
- Textual Entailment
- AAAI ’07 Symposium on “Machine Reading”
- Growing body of work on IE from the Web
Lessons Reprised

- Scale & diversity of the Web $\Rightarrow$ Open IE
- Redundancy helps accuracy!
- Focus on 'tractable' sentences
- Open IE(Web) = disjointed “nuggets”
- Scalable inference is key
- Leverage relations implicit in the text
- Sentence-based IE $\Rightarrow$ page-based IE

Directions for Future Work

- Characterize 'tractable' sentences
- Page-based IE
  - Better entity resolution
- Integrate Open IE with ontologies
- Investigate other domains for Open IE (e.g., news)
Implications for Web Search

Imagine search systems that operate over a (more) semantic space

- Key words, documents $\Rightarrow$ extractions
- TF-IDF, pagerank $\Rightarrow$ entities, relations
- Web pages, links $\Rightarrow$ Q/A, information fusion

Reading the Web $\Rightarrow$ new Search Paradigm

Machine Reading of the Web

Found 2,300 reviews; 85% positive, key features are...

How is the ThinkPad T-40?

Open IE over $10^{10}$ Web pages.
Machine Reading at Web Scale?

- Difficult, ungrammatical sentences
- Tractable Sentences
- Unreliable information
- (PageRank)
- Redundancy
- Unanticipated objects & relations
- Open Information Extraction + Linear-time Inference
- Massive Scale
Holmes - KBMC

KBs
- 0.9 Prevents(Ca, osteo.)
- 0.8 Contains(kale, Ca)
- ...

Inf. Rules
- Weighted Horn Clauses
  - 1.5 P(X,Z) :- C(X,Y) \& P(Y,Z)

Query
- Q(X) :- prevent(X, osteo.)

Find Proof Trees
Construct Markov Network

Answers
Approximate Probabilistic Inference

Chain backwards from the query
Convert proof tree into Markov network

Test

- Assess candidate extraction E using Pointwise Mutual Information (PMI) between E and a “discriminator phrase” D

- PMI-IR (Turney ’01): use hit counts for efficient PMI computation over the Web

\[
PMI(Yakima, City) = \frac{|\text{Hits(Yakima+City)}|}{|\text{Hits(Yakima)}|}
\]
PMI In Action

1. E = “Yakima” (1,340,000)
2. D = <class name>
3. E+D = "Yakima city" (2760)
4. PMI = (2760 / 1.34M)= 0.002

• E = “Avocado” (1,000,000)
• E+D = “Avocado city” (10)
  PMI = 0.00001 << 0.002
Can IE achieve “adequate” precision/recall on the Web corpus?
Recall - Precision Curve
(City, Country, USState)

High Precision for Binary Predicates
Methods for Improving Recall

- **RL:** learn class-specific patterns
  
  "Headquartered in <city>"

- **SE:** Recursively extract subclasses
  
  "Scientists such as physicists, chemists, and biologists…"

- **LE:** extract lists of items & vote

  • Bootstrapped from generic KnowItAll—no hand-labeled examples!

Results for City

Found 10,300 cities missing from Tipster Gazetteer.
Sources of ambiguity

- **Time:** “Bush is the president”
  - in 2006!
- **Context:** “common misconceptions.”
- **Opinion:** Who killed JFK?
- **Multiple word senses:** Amazon, Chicago, Chevy Chase, etc.

Critique of KnowItAll 1.0

- User query (“African cities”) launches search engine queries ➔ slow, limited
- Relations specified by user
- N relations ➔ N passes over the data

Can we make a single pass over the Web corpus?
Extractor Overview (Banko & Etzioni, ‘08)

1. Use a simple model of relationships in English to label extractions
2. Bootstrap a general model of relationships in English sentences, encoded as a CRF
3. Decompose each sentence into one or more (NP1, VP, NP2) “chunks”
4. Use CRF model to retain relevant parts of each NP and VP.

The extractor is relation-independent!
Fundamental AI Questions

- How to accumulate massive amounts of knowledge?

- Can a program read and “understand” what it’s read?
Massive Information Extraction from the Web

How is the ThinkPad T-40?

KnowItAll

Found 2,300 reviews; 85% positive.

Stats over $10^{10}$ web pages.