The Intelligence in Wikipedia Project

Daniel S. Weld
Department of Computer Science & Engineering
University of Washington
Seattle, WA, USA

Joint Work with
Fei Wu, Eytan Adar, Saleema Amershi, Oren Etzioni, James Fogarty, Raphael Hoffmann, Kayur Patel & Stef Schoenmackers
History

MetaCrawler [Selberg & Etzioni WWW]

Wrapper Induction [K.

Shopbot [Doorenbc

YOU KNOW, THIS MIGHT
NOT BE SUCH A CRAZY
IDEA...
History

MetaCrawler [Selberg & Etzioni WWW]
Wrapper Induction [K.
Shopbot [Doorenbc

...EXCEPT THAT OREN THOUGHT
OF IT.
History

MetaCrawler [Selberg & Etzioni WWW]

Wrapper Induction [Dooreh]

Shopbot
History

MetaCrawler [Selberg & Etzioni WWW]
Wrapper Induction [K.
Shopbot [Doorenbo
History

MetaCrawler [Selberg & Etzioni WWW]

Wrapper Induction [Kushmerick et al. IJCAI]

Shopbot [Doorenbos et al. AGENTS]

Data Integration [Ives et al. SIGMOD]

Question Answering [Kwok, Etzioni & Weld WWW]

KnowItAll [Etzioni et al. WWW]

Kylin
Which German Scientists Taught at US Universities?

Albert Einstein was a German-born theoretical physicist...

Einstein was a guest lecturer at the Institute for Advanced Study in New Jersey...

New Jersey is a state in the Northeastern region of the United States...
Next-Generation Search

- Information Extraction
  - \(<\text{Einstein, Born-In, Germany}>\)
  - \(<\text{Einstein, ISA, Physicist}>\)
  - \(<\text{Einstein, Lectured-At, IAS}>\)
  - \(<\text{IAS, In, New-Jersey}>\)
  - \(<\text{New-Jersey, In, United-States}>\)

- Ontology
  - Physicist (x) → Scientist(x)

- Inference
  - Einstein = Einstein
Why Wikipedia?

- Comprehensive
- High Quality
  [Giles Nature 05]
- Useful Structure
  Unique IDs & Links
  Infoboxes
  Categories & Lists
  First Sentence
  Redirection pages
  Disambiguation pages
  Revision History
  Multilingual

Cons

- Natural-Language
- Missing Data
- Inconsistent
- Low Redundancy
The Intelligence in Wikipedia Project

Outline

1. Self-supervised extraction from Wikipedia text (and the greater Web)
2. Automatic ontology generation
3. Scalable probabilistic inference for Q/A
Outline

1. **Self-supervised extraction** from Wikipedia text
2. Automatic **ontology generation**
3. Scalable **probabilistic inference** for Q/A
• Info extraction – defn
• Precision
• Recall
Kirkland-based Microsoft is the largest software company. Boeing moved it’s headquarters to Chicago in 2003. Hank Levy was named chair of Computer Science & Engr. …

HeadquarterOf(<company>,<city>)
Clearfield County was created in 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812. Its county seat is Clearfield. 2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water. As of 2005, the population density was 28.2/km².
Kylin Architecture

Wikipedia

Schema Refiner

Training Data Constructor

Preprocessor

Document Classifier

Sentence Classifier

Classifier

CRF Model

Extractor

Infoboxes
Preliminary Evaluation

- Kylin Performed Well on Popular Classes:
  - Precision: mid 70% ~ high 90%
  - Recall: low 50% ~ mid 90%

- ... Floundered on Sparse Classes – Little Training Data

82% < 100 instances; 40% < 10 instances
Shrinkage?

Diagram:
- Person (1201)
  - Performer (44)
    - Actor (8738)
    - Comedian (106)

Properties:
- birth_place
- location
  - birthplace
  - birth_place
  - city_of_birth
  - origin
Outline

1. **Self-Supervised Extraction** from Wikipedia Text
   - Training on Infoboxes
   - Improving Recall – Shrinkage, Retraining, Web Extraction
   - Community Content Creation

2. **Automatic Ontology Generation**

3. **Scalable Probabilistic Inference** for Q/A
KOG: Kylin Ontology Generator
[Wu & Weld, WWW08]
Subsumption Detection

- Binary Classification Problem
- Nine Complex Features
  - E.g., String Features
  - ... IR Measures
  - ... Mapping to Wordnet
  - ... Hearst Pattern Matches
  - ... Class Transitions in Revision History
- Learning Algorithm
  SVM & MLN Joint Inference
KOG Architecture
Schema Mapping

- **Heuristics**
  - Edit History
  - String Similarity

- **Experiments**
  - Precision: 94%  Recall: 87%

- **Future**
  - Integrated Joint Inference
Outline

1. **Self-Supervised Extraction** from Wikipedia Text
   - Training on Infoboxes
   - **Improving Recall** – Shrinkage, Retraining, Web Extraction
   - Community Content Creation

2. Automatic **Ontology Generation**

3. Scalable **Probabilistic Inference** for Q/A
Improving Recall on Sparse Classes

[Wu et al. KDD-08]

- **Shrinkage**
  - Extra Training Examples from Related Classes
  - How Weight New Examples?

- **Retraining**
  - Compare Kylin Extractions with Ones from Textrunner [Banko et al. IJCAI-07]
  - Additional Positive Examples
  - Eliminate False Negatives

- **Extraction from Broader Web**
Effect of Shrinkage & Retraining

1755% improvement
for a sparse class

13.7% improvement
for a popular class
Improving Recall on Sparse Classes

- Shrinkage
- Retraining

- Extract from Broader Web
  - 44% of Wikipedia Pages = “stub”
    - Extractor quality irrelevant
  - Query Google & Extract
    - How maintain high precision?
    - Many Web pages noisy, describe multiple objects
    - How integrate with Wikipedia extractions?
Recall after Shrinkage / Retraining...
Combining Wikipedia & Web

(a) Irish Newspaper

(b) Performer

(c) Baseball Stadium

(d) Writer
Outline

1. **Self-Supervised Extraction** from Wikipedia Text
   - Training on Infoboxes
   - Improving Recall – Shrinkage, Retraining, Web Extraction
   - Community Content Creation

2. Automatic **Ontology Generation**

3. Scalable **Probabilistic Inference** for Q/A
Problem

- Information Extraction is Imprecise
  - Wikipedians Don’t Want 90% Precision

- How Improve Precision?
  - People!
Contributing as a Non-Primary Task

- Encourage contributions
- Without annoying or abusing readers
  - Compared 5 different interfaces
Ray Bradbury

From Wikipedia, the free encyclopedia

Ray Douglas Bradbury (born August 22, 1920) is an American literary, fantasy, horror, science fiction, and mystery writer best known for *The Martian Chronicles*, a 1950 book which has been described both as a short story collection and a novel, and his 1953 dystopian novel *Fahrenheit 451*. He is widely considered to be one of the greatest and most popular American writers of speculative fiction during the twentieth century.

### Beginnings

Bradbury was born in Waukegan, Illinois, to a Swedish immigrant mother and a father who was a power and telephone company lineman. His paternal grandfather departed Sweden with his nursing mother only months after arriving in New York, spending two years in a New York orphanage while his mother, a Waukegan seamstress, searched for him. born in Waukegan, his name is the same as the town and is also the birthplace of the book *Dandelion Wine*, one of his other semi-autobiographical novels. He graduated from Waukegan's Groebe High School and 1944 in *Farewell Summer* — as well as in many of his short stories.
Adwords Deployment Study

- 2000 articles containing writer infobox
- Query for “ray bradbury” would show

Ray Bradbury - Wikipedia
Get enhanced Wikipedia content for Ray Bradbury.
intelligent-wikipedia.org

- Redirect to mirror with injected JavaScript
- Round-robin interface selection:
  - baseline, popup, highlight, icon
- Track clicks, load, unload, and show survey

[Hoffman et al. 2008]
Results

- Contribution Rate
  - 1.6% → 13%
- 90% of positive labels were correct
Outline

1. Self-Supervised Extraction from Wikipedia Text

2. Automatic Ontology Generation

3. Scalable Probabilistic Inference for Q/A
Scalable Probabilistic Inference

- Eight MLN Inference Rules
  - Transitivity of predicates, *etc.*
- Knowledge-Based Model Construction
- Tested on 100 Million Tuples
  - Extracted by Textrunner from Web

[Schoenmacker *et al.* 2008]
Effect of Limited Inference

The graph shows the precision at various estimated recall levels for two methods: Baseline and Holmes. The Baseline method starts with a higher precision but drops earlier compared to the Holmes method, which maintains a higher precision even at higher estimated recall levels. The area under the curve (AuC) increases significantly for the Holmes method, indicating a better performance in terms of precision and recall trade-off.

Increase in AuC

- **Baseline**
- **Holmes**
Cost of Inference

Approximately Pseudo-Functional Relations
Related Work

- **Unsupervised Information Extraction**
  - SNOWBALL [Agichtein & Gravano ICDL00]
  - MULDER [Kwok et al. TOIS01]
  - AskMSR [Brill et al. EMNLP02]
  - KnowItAll [Etzioni et al. WWW04]
  - TextRunner [Banko et al. IJCAI07]

- **Ontology Driven Information Extraction**
  - SemTag and Seeker [Dill WWW03]
  - PANKOW [Cimiano WWW05]
  - OntoSyphon [McDowell & Cafarella ISWC06]
Related Work II

- **Other Uses of Wikipedia**
  - semantic distance measure [Ponzetto & Strube 07]
  - Word-Sense Disambiguation [Bunescu & Pasca 06, Mihalcea 07]
  - Coreference Resolution [Ponzetto & Strube 06, Yang & Su 07]
  - Ontology / Taxonomy [Suchanek 07, Muchnik 07]
  - Multi-Lingual Alignment [Adafre & Rijke 06]
  - Question Answering [Ahn et al. 05, Kaisser 08]
  - Basis of Huge KB [Auer et al. 07]
Conclusion

- **Wikipedia is a Fantastic Platform & Corpus**

- **Self-Supervised Extraction** from Wikipedia
  Training on Infoboxes
  - Works well on popular classes
  Improving Recall – Shrinkage, Retraining, Web Extraction
  - High precision & recall - even on sparse classes, stub articles

- **Community Content Creation**

- **Automatic Ontology Generation**
  Probabilistic Joint Inference

- **Scalable Probabilistic Inference** for Q/A
  Simple Inference - Scales to Large Corpora
  Tested on 100 M Tuples
Future Work

- Improve Precision
- Automatically Learn Inference Rules
- Multi-Lingual Extraction
- Improved Ontology Generation
  - Joint Schema Mapping
  - Incorporate Freebase, etc.
- Make Available as Web Service
  - Integrate Back Into Wikipedia
Can we create better features?

Example:

training data
A.B. is a **Swedish** politician.
Y.P. was a **Russian** numismatist.
J.F.S.d.L. is a **Brazilian** attacking midfielder.
B.C.A. is a **Danish** former football player.

test data
N.F. was an *Australian*.
R.S. is a *Fijian* politician.
A. I. is a Boeing subsidiary.
D.F. is a *Swedish* actor.

now:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>isCapitalized</td>
<td>1</td>
</tr>
<tr>
<td>contains n-gram “ed”</td>
<td>0</td>
</tr>
<tr>
<td>equals “Swedish”</td>
<td>0</td>
</tr>
<tr>
<td>equals “Russian”</td>
<td>1</td>
</tr>
<tr>
<td>equals “politician”</td>
<td>0</td>
</tr>
<tr>
<td>equals “B.C.A.”</td>
<td>0</td>
</tr>
</tbody>
</table>

would like:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>is on list of occupations</td>
<td>0</td>
</tr>
<tr>
<td>is on list of nationalities</td>
<td>1</td>
</tr>
<tr>
<td>is on list of first names</td>
<td>0</td>
</tr>
<tr>
<td>is on list of other</td>
<td>0</td>
</tr>
</tbody>
</table>

But where do we get the lists from?
Mining lists from the Web

A.B. is a **Swedish** politician.
Y.P. was a **Russian** numismatist.
J.F.S.d.L. is a **Brazilian** attacking midfielder.
B.C.A. is a **Danish** former football player.

55 million lists

unified list
Tom was born in Seattle.

<table>
<thead>
<tr>
<th>Sub</th>
<th>Pred</th>
<th>Obj</th>
<th>Snt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>BirthPlace</td>
<td>Seattle</td>
<td></td>
</tr>
</tbody>
</table>

**Tom Lau**

<table>
<thead>
<tr>
<th>Sub</th>
<th>Pred</th>
<th>Obj</th>
<th>Snt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom Lau</td>
<td>BirthPlace</td>
<td>Seattle, WA</td>
<td></td>
</tr>
</tbody>
</table>
Many Thanks To:

Eytan Adar
Saleema Amershi
Oren Etzioni
James Fogarty
Raphael Hoffmann
Kayur Patel
Stef Schoenmackers
Fei Wu
Thanks!

Fei Wu, Eytan Adar, Saleema Amershi, Oren Etzioni, James Fogarty, Raphael Hoffmann, Kayur Patel & Stef Schoenmackers
Next-Generation Search = Information Extraction + Ontology

KOG
[Wu & Weld WWW08]
## Web Extractions Experiment

### Improvement of Shrinkage-Retrain-Web over baseline

<table>
<thead>
<tr>
<th></th>
<th>Irish newspaper</th>
<th>Performer</th>
<th>Baseball stadium</th>
<th>Writer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>742.9%</td>
<td>89.9%</td>
<td>56.3%</td>
<td>76.9%</td>
</tr>
<tr>
<td>Area under PR curve</td>
<td>1771.3%</td>
<td>101.8%</td>
<td>90.6%</td>
<td>93.5%</td>
</tr>
</tbody>
</table>

### Improvement of Shrinkage-Retrain over baseline

<table>
<thead>
<tr>
<th></th>
<th>Irish newspaper</th>
<th>Performer</th>
<th>Baseball stadium</th>
<th>Writer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>585.3%</td>
<td>73.2%</td>
<td>29.8%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Area under PR curve</td>
<td>1755%</td>
<td>78.9%</td>
<td>42.5%</td>
<td>13.7%</td>
</tr>
<tr>
<td>AUC improved(%)</td>
<td>+Shrink</td>
<td>+Retrain</td>
<td>+Web</td>
<td>Total</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>Irish news.(20)</td>
<td>1386</td>
<td>294</td>
<td>552</td>
<td>2232</td>
</tr>
<tr>
<td>Performer(44)</td>
<td>57</td>
<td>17</td>
<td>24</td>
<td>98</td>
</tr>
<tr>
<td>Baseball stad.(163)</td>
<td>17</td>
<td>23</td>
<td>62</td>
<td>102</td>
</tr>
<tr>
<td>Writer(2213)</td>
<td>7</td>
<td>9</td>
<td>80</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 2: Accumulative AUC improvements.
Intelligence in Wikipedia Project

Textrunner Retraining

Web Extraction

MultiLingual Extraction

Preprocessor

Document & Sentence Classifier

Shrinkage

CRF Extractor

Infoboxes