Machine Reading
From Wikipedia to the Web

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

todo
- More on bootstrapping to the web
- Retrain too brief
- Results for shrinkage independent of retraining

Many Collaborators...

Raphael Hoffmann
Stefan Schoenmackers
Fei Wu

And… Eytan Adar, Saleema Amershi, Oren Etzioni,
James Fogarty, Xiao Ling, Kayur Patel

Overview

- Extracting Knowledge from the Web
  - Facts
  - Ontology
  - Inference Rules
- Using it for Q/A

Key Ideas

Key Idea 1  Ways WWW → Knowledge

Machine-Learning-Based Information Extraction

Community Content Creation

UW Intelligence in Wikipedia Project
**Key Idea 1**
- **Synergy (Positive Feedback)**
  - Between ML Extraction & Community Content Creation

**Key Idea 2**
- **Synergy (Positive Feedback)**
  - Between ML Extraction & Community Content Creation
- **Self Supervised Learning**
  - Heuristics for Generating (Noisy) Training Data

**Key Idea 3**
- **Synergy (Positive Feedback)**
  - Between ML Extraction & Community Content Creation
- **Self Supervised Learning**
  - Heuristics for Generating (Noisy) Training Data
- **Shrinkage (Ontological Smoothing) & Retraining**
  - For Improving Extraction in Sparse Domains

**Key Idea 4**
- **Synergy (Positive Feedback)**
  - Between ML Extraction & Community Content Creation
- **Self Supervised Learning**
  - Heuristics for Generating (Noisy) Training Data
- **Shrinkage (Ontological Smoothing) & Retraining**
  - For Improving Extraction in Sparse Domains
- **Approximately Pseudo-Functional (APF) Relations**
  - Efficient Inference Using Learned Rules

**Motivating Vision**
Next-Generation Search = Information Extraction + Ontology + Inference
- Which German Scientists Taught at US Universities?

**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) \(\Rightarrow\) Scientist(x)
- **Inference**
  - Lectured-At(x, y) \& University(y) \(\Rightarrow\) Taught-At(x, y)
  - Einstein = Einstein
Open Information Extraction

<table>
<thead>
<tr>
<th>Table 2: The contrast between traditional and open IE.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traditional IE</strong></td>
</tr>
<tr>
<td>Input</td>
</tr>
<tr>
<td>Relations</td>
</tr>
<tr>
<td>Complexity</td>
</tr>
</tbody>
</table>

TextRunner

For each sentence
Apply POS Tagger
For each pairs of noun phrases, NP₁, NP₂
If classifier confirms they are “Related?”
Use CRF to extract relation from intervening text
Return relation(NP₁, NP₂)

Train classifier & extractor on Penn Treebank data

Mark Emmert was-born-in Fife and graduated from UW in 1975

Why Wikipedia?
- **Pros**
  - Comprehensive
  - High Quality
    [Giles Nature 05]
  - Useful Structure
- **Cons**
  - Natural-Language
  - Missing Data
  - Inconsistent
  - Low Redundancy

Wikipedia Structure
- Unique IDs & Links
- Infoboxes
- Categories & Lists
- First Sentence
- Redirection pages
- Disambiguation pages
- Revision History
- Multilingual
Status Update

Outline
- Motivation
- Extracting Facts from Wikipedia
- Ontology Generation
- Improving Fact Extraction
- Bootstrapping to the Web
- Validating Extractions
- Improving Recall with Inference
- Conclusions

Key Ideas
- Synergy
- Self-Supervised Learning
- Extraction Generation
- Improving Fact Extraction
- Bootstrapping to the Web
- Validating Extractions
- Improving Recall with Inference
- Conclusions

Traditional, Supervised I.E.

Raw Data

Labeled Training Data

Learning Algorithm

Extractor

Kylin: Self-Supervised Information Extraction from Wikipedia

From infoboxes to a training set

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.

The Precision / Recall Tradeoff

- Precision \[ \frac{tp}{tp + fp} \]
  - Proportion of selected items that are correct
- Recall \[ \frac{tp}{tp + fn} \]
  - Proportion of target items that were selected
- Precision-Recall curve
  - Shows tradeoff

Kylin Architecture

Kylin: Self-Supervised Information Extraction from Wikipedia

From infoboxes to a training set

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.

The Precision / Recall Tradeoff

- Precision \[ \frac{tp}{tp + fp} \]
  - Proportion of selected items that are correct
- Recall \[ \frac{tp}{tp + fn} \]
  - Proportion of target items that were selected
- Precision-Recall curve
  - Shows tradeoff

Kylin Architecture

Kylin: Self-Supervised Information Extraction from Wikipedia

From infoboxes to a training set

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.

The Precision / Recall Tradeoff

- Precision \[ \frac{tp}{tp + fp} \]
  - Proportion of selected items that are correct
- Recall \[ \frac{tp}{tp + fn} \]
  - Proportion of target items that were selected
- Precision-Recall curve
  - Shows tradeoff

Kylin Architecture

Kylin: Self-Supervised Information Extraction from Wikipedia

From infoboxes to a training set

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.

The Precision / Recall Tradeoff

- Precision \[ \frac{tp}{tp + fp} \]
  - Proportion of selected items that are correct
- Recall \[ \frac{tp}{tp + fn} \]
  - Proportion of target items that were selected
- Precision-Recall curve
  - Shows tradeoff

Preliminary Evaluation

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

... But Floundered on Sparse Classes
(Too Little Training Data)

Is this a Big Problem?
Long Tail: Sparse Classes

Too Little Training Data

82% < 100 instances; 40% < 10 instances

Long-Tail 2: Incomplete Articles

- Desired Information Missing from Wikipedia
  800,000/1,800,000 (44.2%) stub pages [Wikipedia July 2007]

Shrinkage?

Status Update

Outline
- Motivation
- Extracting Facts from Wikipedia
- Ontology Generation
- Improving Fact Extraction
- Bootstrapping to the Web
- Validating Extractions
- Improving Recall with Inference
- Conclusions

Key Ideas
- Synergy
- Self-Supervised Learning
- Shrinkage & Retraining
- APF Relations

KOG: Kylin Ontology Generator

[Wu & Weld, WWW08]

How Can We Get a Taxonomy for Wikipedia?

Do We Need to?
- What about Category Tags?
  - Conjunctions
  - Schema Mapping

Person

Performer

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

KOG: Kylin Ontology Generator

Status Update

Outline
- Motivation
- Extracting Facts from Wikipedia
- Ontology Generation
- Improving Fact Extraction
  - Bootstrapping to the Web
  - Validating Extractions
  - Improving Recall with Inference
- Conclusions

Key Ideas
- Synergy
- Self-Supervised Learning
- Shrinkage & Retraining
- APF Relations

Improving Recall on Sparse Classes

- Shrinkage
  - Extra Training Examples from Related Classes
  - How Weight New Examples?
Improving Recall on Sparse Classes

[Wu et al. KDD-08]

Retraining
- Compare Kylin Extractions with Tuples from Textrunner
- Additional Positive Examples
- Eliminate False Negatives

TextRunner [Banko et al. IJCAI-07, ACL-08]
- Relation-Independent Extractions
- Exploits Grammatical Structure
- CRF Extractor with POS Tag Features

Status Update

Outline
- Motivation
- Extracting Facts from Wikipedia
- Ontology Generation
- Improving Fact Extraction
- Bootstrapping to the Web
- Validating Extractions
- Improving Recall with Inference
- Conclusions

Key Ideas
- Synergy
- Self-Supervised Learning
- Shrinkage & Retraining
- APF Relations

Bootstrapping to the Web

[Wu et al. KDD-08]

- Extractor Quality Irrelevant
  - If no information to extract...
  - 44% of Wikipedia Pages = “stub”

- Instead, ... Extract from Broader Web

- Challenges
  - How maintain high precision?
  - Many Web pages noisy,
  - Describe multiple objects

Long-Tail 2: Incomplete Articles

- Desired Information Missing from Wikipedia
  - 800,000/1,800,000 (44.2%) stub pages [July 2007 of Wikipedia]

Extracting from the Broader Web

1) Send Query to Google
   Object Name + Attribute Synonym

2) Find Best Region on the Page
   Heuristics > Dependency Parse

3) Apply Extractor

4) Vote if Multiple Extractions
Problem

- Information Extraction is Still Imprecise
  - Do Wikipedians Want 90% Precision?
- How Improve Precision?
  - People!

Status Update

Outline
- Motivation
- Extracting Facts from Wikipedia
- Ontology Generation
- Improving Fact Extraction
- Bootstrapping to the Web

Validating Extractions
- Improving Recall with Inference

Key Ideas

Synergy
- Self-Supervised Learning
- Shrinkage & Retraining
- APF Relations

Contributing as a Non-Primary Task

- Encourage contributions
- Without annoying or abusing readers

Designed Three Interfaces

- Popup
  (immediate interruption strategy)
- Highlight
  (negotiated interruption strategy)
- Icon
  (negotiated interruption strategy)
How do you evaluate these UIs?

*Contribution as a non-primary task*

Can lab study show if interfaces increase *spontaneous* contributions?

**Search Advertising Study**

- Deployed interfaces on Wikipedia proxy
- 2000 articles
- One ad per article

**Search Advertising Study**

- Select interface round-robin
- Track session ID, time, all interactions
- Questionnaire pops up 60 sec after page loads

- Used Yahoo and Google
- Deployment for ~7 days
  - ~1M impressions
  - 2473 visitors
**Contribution Rate > 8x**

**Area under Precision/Recall curve with only existing infoboxes**

**Area under Precision/Recall curve after adding user contributions**

**Search Advertising Study**
- Used Yahoo and Google
- 2473 visitors
- Estimated cost: $1500
  
  Hence ~$10 / contribution !!

**Status Update**

**Key Ideas**
- Synergy
- Self-Supervised Learning
- Shrinkage & Retraining
- APF Relations

**Why Need Inference?**

- **What Vegetables Prevent Osteoporosis?**

**No Web Page Explicitly Says:**

"Kale is a vegetable which prevents Osteoporosis"

But some say
- "Kale is a vegetable" …
- "Kale contains calcium" …
- "Calcium prevents osteoporosis"
Three Part Program

1) Scalable Inference with Hand Rules
   In small domains (5-10 entity classes)

2) Learning Rules for Small Domains

3) Scaling Learning to Larger Domains
   E.g., 200 entity classes

---

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

---

Effect of Limited Inference

Inference Appears Linear in |Corpus|

- $Q(X,Y,Z) \Leftarrow \text{Married}(X,Y) \land \text{LivedIn}(Y,Z)$
- Worst Case: Some person $y'$ married everyone, and lived in every place:
  $|Q(X,y',Z)| = |\text{Married}| \cdot |\text{LivedIn}| = O(n^2)$

---

How Can This Be True?

What makes inference expensive?

- $Q(X,Y,Z) \Leftarrow \text{Married}(X,Y) \land \text{LivedIn}(Y,Z)$
- Common Case: Essentially functional
  A few spouses and a few locations.

- Person
  Ramesses II (100+).
  Elizabeth Taylor (7).
### Approximately Pseudo-Functional Relations

E.g. Married(X,Y) Most Y have only 1 spouse mentioned

People in \( Y \) have at most a constant \( k_m \) spouses each

People in \( X \) have at most \( k_m \log |X| \) spouses in total

![Theorem](image)

### Prevalence of APF relations

![Graph showing the prevalence of APF relations](image)

### Learning Rules

- **Work in Progress**
  - Tight Bias on Rule Templates
    - Entailment: \( R_1(X,Y) \rightarrow R_2(X,Y) \)
    - Homophily: \( R_1(X,Y) \rightarrow R_2(X,Z) \land R_2(Y,Z) \)
    - Generalized transitivity: \( R_1(X,Z) \land R_2(Y,Z) \rightarrow R_2(X,Y) \)
  - Type Constraints on Shared Variables
  - Mechanical Turk Validation
    - 20% → 90% precision
  - **Learned Rules Beat Hand-Coded**
    - On small domains
    - **Now Scaling to 200 Entity Classes**

### Status Update

<table>
<thead>
<tr>
<th>Outline</th>
<th>Key Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td>Synergy</td>
</tr>
<tr>
<td>Extracting Facts from Wikipedia</td>
<td>Self-Supervised Learning</td>
</tr>
<tr>
<td>Ontology Generation</td>
<td>Shrinkage &amp; Retraining</td>
</tr>
<tr>
<td>Improving Fact Extraction</td>
<td>APF Relations</td>
</tr>
<tr>
<td>Bootstrapping to the Web</td>
<td></td>
</tr>
<tr>
<td>Validating Extractions</td>
<td></td>
</tr>
<tr>
<td>Improving Recall with Inference</td>
<td></td>
</tr>
</tbody>
</table>

### Conclusion

- **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
Conclusion

- **Extraction of Facts** from Wikipedia & Web
  Self-Supervised Training on Infoboxes
  Improving Recall – Shrinkage, Retraining,
  Need for Humans to Validate

- **Automatic Ontology Generation**
  Probabilistic Joint Inference

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

Key Ideas

- **Synergy (Positive Feedback)**
  - Between ML Extraction & Community Content Creation

- **Self Supervised Learning**
  - Heuristics for Generating (Noisy) Training Data

- **Shrinkage & Retraining**
  - For Improving Extraction in Sparse Domains

- **Aproximately Pseudo-Functional Relations**
  - Efficient Inference Using Learned Rules

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. UCAI07, ACL-08]
  - KNEXT [VanDurme et al. COLING-08]
  - WebTables [Cafarella et al. VLDB-08]

- **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&Strube07]
  - Word-Sense Disambiguation [Bunescu&Pasca06, Mihalcea07]
  - Coreference Resolution [Ponzetto&Strube06, Yang&Su07]
  - Ontology / Taxonomy [Suchanek07, Muchnik07]
  - Multi-Lingual Alignment [Adafre&Rijke06]
  - Question Answering [Ahn et al.05, Kaisser08]
  - Basis of Huge KB [Auer et al.07]

Thanks!

In Collaboration with

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

Funding Support

NSF, ONR, DARPA, WRF TJ Cable Professorship,
Google, Yahoo