Autonomously
Semantifying Wikipedia

Fei Wu
University of Washington

(Joint work with Dan Weld)
Motivation (1)

- Semantic Web [Berners-Lee 01] is great.
  - Web content machine readable
  - Software agents find, share and integrate information
Motivation (1)

- Semantic Web [Berners-Lee 01] is great.
  - Web content machine readable
  - Software agents find, share and integrate information

Chicken-egg problem:
Motivation (1)

- Semantic Web [Berners-Lee 01]
  - Web content machine readable
  - Software agents find, share and integrate information

Chicken-egg problem:

Bootstrapping:

Automatically Semantifying Data
Motivation (2)

- Wikipedia [http://wikipedia.org]
  - Comprehensive (1.7 million English articles)
  - High-quality
  - User-tagged data (links, infobox, lists, categories, etc.)

Challenges:

- Missing data
- Inconsistency
- Low information redundancy
Kylin: Autonomously Semantifying Wikipedia

✓ Totally autonomous with no additional human efforts
✓ Information extraction from both semi-structured and unstructured data

Kylin: a *mythical* hooved *Chinese* *chimerical* creature that is said to appear in conjunction with the arrival of a sage.

------ Wikipedia
Outline

● Semantics in Wikipedia
  ▪ Opportunities
  ▪ Challenges
● Kylin System
  ▪ Infobbox Generation
  ▪ Link Creation
● Conclusion
Semantics in Wikipedia

- **Infobox**
- **Link**
- **List**
- **Category**
- **Redirection**
- **Disambiguation**

```
{{Infobox U.S. County
  county  = Clearfield County |
  state = Pennsylvania |
  seal = |
  map = Map of Pennsylvania
    highlighting Clearfield County.svg |
  map size = 225 |
  founded = [[March 26]], [[1804]] |
  seat = [[Clearfield, Pennsylvania|Clearfield]] |
  area = 2,988 [[km²]] (1,154 [[square mile(mi²)]] |
  area water = 17 km² (6 mi²) |
  area percentage = 0.56% |
  census yr = 2000 |
  pop = 83,382 |
  density = 28 |
}}
```
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

Seattle, Washington

From Wikipedia, the free encyclopedia

“Seattle” redirects here. For other uses, see Seattle (disambiguation).

Seattle is the largest city in the Pacific Northwest region of the United States. It is in the U.S. state of Washington between Puget Sound and Lake Washington, nearly 108 miles south of the United States–Canadian border in King County, of which it is the county seat.

Seattle was first settled November 14, 1851, by Arthur Denny, which subsequently become known as the Denny party. Its first name was New York, then finally it was renamed Seattle after Chief Noah Sealth who was chief of the two.
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

List of cities in Washington

From Wikipedia, the free encyclopedia

This is a list of cities in Washington, U.S.A. See also List of cities in Washington (by population).

Contents: Top · 0-9 · A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S

List of cities in California

From Wikipedia, the free encyclopedia

There are 479 incorporated cities in California, 22 of which are styled “Town.” (Government Code Sections 34500-34504) the terms “city” and “town” are defined:

Contents: Top · 0-9 · A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z

<table>
<thead>
<tr>
<th>City</th>
<th>County</th>
<th>Incorporated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelanto</td>
<td>San Bernardino</td>
<td>December 22, 1970</td>
</tr>
<tr>
<td>Agoura Hills</td>
<td>Los Angeles</td>
<td>December 8, 1982</td>
</tr>
<tr>
<td>Alameda</td>
<td>Alameda</td>
<td>April 19, 1854</td>
</tr>
<tr>
<td>Albany</td>
<td>Alameda</td>
<td>September 22, 1908</td>
</tr>
</tbody>
</table>
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

“Seattle, Washington”

Categories: Move protected | Wikipedia articles needing factual verification | Articles with unsourced statements since January 2007 | All articles with unsourced statements | Articles with unsourced statements since May 2007 | Articles with unsourced statements since February 2007 | Articles with unsourced statements since April 2007 | 1850s establishments | Seattle, Washington | King County, Washington | Cities in the Seattle metropolitan area | County seats in Washington | Towns and cities with limited zero-fare transport
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

America

From Wikipedia, the free encyclopedia

America usually means either:
- The Americas, the lands and regions of the Western hemisphere, often divided into North America and South America
- The United States of America

Other places

- America, Netherlands
- América, Buenos Aires in Argentina
- 916 America, a minor planet of the solar system
- America, a small town in Italy, during the time of the Roman Republic. It is the home town of Sextus Roscius, who was accused of parricide and defended by Cicero.

Art and entertainment

- America (band), a rock and roll band
- America (XM), XM Satellite Radio channel 13
- America (video game), a real-time strategy video game
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

Opportunities
- Semantic source
- Training dataset

Challenges
- Missing data
- Inconsistency
Semantics in Wikipedia

- Infobox
- Link
- List
- Category
- Redirection
- Disambiguation

Opportunities
- Semantic source
- Training dataset

Challenges
- Missing data
- Inconsistency

Kylin: Autonomously Semantifying Wikipedia
Outline

● Semantics in Wikipedia
  ▪ Opportunities
  ▪ Challenges

● Kylin System
  ▪ Infobox Generation
  ▪ Link Creation

● Conclusion
Infobox Generation
Preprocessor

- Schema Refinement
  - Free edit -> schema drift
  - Duplicate templates:
    U.S.County(1428), US County(574), Counties(50), County(19)
  - Duplicate attributes:
    “Census Yr”, “Census Estimate Yr”, “Census Est.”, “Census Year”
  - Low usage of attribute

Kylin:
- Strict name match
- >15% occurrences
Preprocessor

- Training Dataset Construction

<table>
<thead>
<tr>
<th>Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Founded</strong></td>
<td>March 26, 1804</td>
</tr>
<tr>
<td><strong>Seat</strong></td>
<td>Clearfield</td>
</tr>
<tr>
<td><strong>Area</strong></td>
<td></td>
</tr>
<tr>
<td>- Total</td>
<td>2,988 km² (1,154 mi²)</td>
</tr>
<tr>
<td>- Land</td>
<td>sq mi (km²)</td>
</tr>
<tr>
<td>- Water</td>
<td>17 km² (6 mi²), 0.56%</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td></td>
</tr>
<tr>
<td>- (2000)</td>
<td>83,382</td>
</tr>
<tr>
<td>- Density</td>
<td>28/km²</td>
</tr>
</tbody>
</table>

Clearfield County was created on 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.

Problems:

- Missing data
- Noise
Classifier

- Document Classifier
  - List & Category
    - Fast
    - Precision(98.5%)
    - Recall(68.8%)

- Sentence Classifier
  - Maximum Entropy Classifier with Bagging: multi-class, multi-label, missing data
CRF Extractor

- Conditional Random Fields Model [Lafferty 01]
  Attribute value extraction: sequential data labeling
  - CRF model for each attribute independently

- Relabel — filter false negative training examples

  2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

  Preprocessor: Water_area

  Classifier: Water_area; Land_area

- Pipeline — prune irrelevant sentences

  Precision + Recall -
Infobox Generation Experiments

- **Dataset**
  - 2007.02.06 Wikipedia Dump Data
    - 4 popular classes:
      - U.S.County(1245)  Actor(3819)
      - Airline(791)  University(4025)
    - 50 random test articles per class
Kylin performance
Abbeville County is a county located in the U.S. state of South Carolina.

The county has a total area of 2,988 square kilometers (1,154 mi²). 2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.
Kylin performance (detailed view)

- University (worse than manual labeling)
  - Flexible expression:
    - The College **began first in 1855** as a one room schoolhouse.
    - **UCL was founded in 1826** under the name “University of London”.
    - The college **opened in 1973** with the Charlestown campus.
  - Global context:
    - Former U.S. President **Dwight D. Eisenhower** served as President of the University.
  - Implicit:
    - Eg: students at 3 campus sum up to the total student number
Effect of Relabel, Pipeline
Outline

● Semantics in Wikipedia
  ▪ Opportunities
  ▪ Challenges

● Kylin System
  ▪ Infobox Generation
  ▪ Link Creation

● Conclusion
Automatic Link Generation

- Link – disambiguation reference

Seattle is located between Puget Sound, an inlet of the Pacific Ocean, and Lake Washington. West beyond the Sound are the Olympic Mountains; Puget Sound

- Kylin: heuristic

<table>
<thead>
<tr>
<th>Proper NP (1st token capitalized)</th>
<th>MatchTitle</th>
<th>MatchAnchor</th>
<th>MatchURI</th>
<th>Disambiguation</th>
<th>InTitle</th>
<th>InAnchor</th>
</tr>
</thead>
</table>

Link Creation Experiments

- 50 random pages from Wikipedia

Training set: 851 existing links of proper NPs
Testing set: 369 more proper NPs missing links

- Human: Pre. ~100% (presumably)
  Rec. 69.8%
- Kylin: Pre. 89.7%
  Rec. 70.7%

\[
\text{Human: Pre. } \sim 100\% \quad (\text{presumably})
\]
\[
\text{Human: Rec. } 69.8\%
\]
\[
\text{Kylin: Pre. } 89.7\%
\]
\[
\text{Kylin: Rec. } 70.7\%
\]
Conclusion

- Semantic Web – Kylin: bootstrapping way

- Wikipedia: great resource for bootstrapping
  - Opportunities: comprehensive, high quality, user-tagged data
  - Challenges: missing data, inconsistency, low redundancy

- Kylin: autonomously semantifying Wikipedia
  - Infobox generation
  - Internal link creation
  - Cope with missing data
Future Work

- Schema Refinement
  - Schema Matching from DB community
- Information Verifier
  - Self-verify accuracy by leveraging outside Web
- Taxonomy Construction
  - Lists, categories, etc.
- Topic Discovery
  - Extend Wikipedia’s coverage
Related Work

- **Bootstrapping Semantic Web**
  - Revere [Halevy 03]
  - DeepMiner[Doan 05]

- **Unsupervised Web IE**
  - AskMSR[Dumais 02]
  - KnowItAll[Etzioni 05], TextRunner[Banko 07]

- **IE from Wikipedia**
  - DBpedia[Auer 07]
  - Yago[Suchanek 07]
  - [Nguyen 07]

- **Related Wikipedia-based System**
  - Semantic Wikipedia[Volkel 06]