Information Extraction

(~20 slides from E. Agichtein)

Steffen Staab
Maciej Janik

Semantic Web
2009-07-10
The Value of Text Data

- “Unstructured” text data is the **primary** form of human-generated information
  - Blogs, web pages, news, scientific literature, online reviews, …
  - The techniques discussed here are complimentary to structured object extraction methods

- Need to extract **structured** information to effectively manage, search, and mine the data
We have a new invention. It is called the Semantic Web!
The annotation problem from a scientific point of view

It will be successful if everybody annotates his/her webpage.
So, we need to annotate all our pages!

That sounds like hard work!

Annotating?
The vicious cycle

Sorry boss, but we have an important appointment...

LET'S RUN!!!
Information Extraction: mature, but active research area
- Intersection of Computational Linguistics, Machine Learning, Data mining, Databases, and Information Retrieval
- Traditional focus on accuracy of extraction
For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

```
<table>
<thead>
<tr>
<th>Name</th>
<th>Title</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bill Gates</td>
<td>CEO</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Bill Veghte</td>
<td>VP</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Richard Stallman</td>
<td>Founder</td>
<td>Free Soft...</td>
</tr>
</tbody>
</table>
```

(dp from William Cohen’s IE tutorial, 2003)
Outline

- Information Extraction Tasks
  - Entity tagging
  - Relation extraction
  - Event extraction

- Scaling up Information Extraction
  - Focus on scaling up to large collections (where data mining can be most beneficial)
  - Other dimensions of scalability
Information Extraction Tasks

- Extracting entities and relations: this talk
  - Entities: named (e.g., Person) and generic (e.g., disease name)
  - Relations: entities related in a predefined way (e.g., Location of a Disease outbreak, or a CEO of a Company)
  - Events: can be composed from multiple relation tuples

- Common extraction subtasks:
  - Preprocess: sentence chunking, syntactic parsing, morphological analysis
  - Create rules or extraction patterns: hand-coded, machine learning, and hybrid
  - Apply extraction patterns or rules to extract new information
  - Postprocess and integrate information
    - Co-reference resolution, deduplication, disambiguation
Entity Tagging

- Identifying mentions of entities (e.g., person names, locations, companies) in text
  - MUC (1997): Person, Location, Organization, Date/Time/Currency
  - ACE (2005): more than 100 more specific types

- Hand-coded vs. Machine Learning approaches

- Best approach depends on entity type and domain:
  - **Closed class** (e.g., geographical locations, disease names, gene & protein names): hand coded + dictionaries
  - **Syntactic** (e.g., phone numbers, zip codes): regular expressions
  - **Semantic** (e.g., person and company names): mixture of context, syntactic features, dictionaries, heuristics, etc.
  - “Almost solved” for common/typical entity types
Example: Extracting Entities from Text

- Useful for data warehousing, data cleaning, web data integration

**Address**

<table>
<thead>
<tr>
<th>House number</th>
<th>Building</th>
<th>Road</th>
<th>City</th>
<th>State</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>4089</td>
<td>Whispering Pines</td>
<td>Nobel Drive</td>
<td>San Diego</td>
<td>CA</td>
<td>92122</td>
</tr>
</tbody>
</table>

**Citation**

Ronald Fagin, *Combining Fuzzy Information from Multiple Systems*, *Proc. of ACM SIGMOD*, 2002

<table>
<thead>
<tr>
<th>Segment($s_i$)</th>
<th>Sequence</th>
<th>Label($s_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>Ronald Fagin</td>
<td>Author</td>
</tr>
<tr>
<td>$S_2$</td>
<td>Combining Fuzzy Information from Multiple Systems</td>
<td>Title</td>
</tr>
<tr>
<td>$S_3$</td>
<td><em>Proc. of ACM SIGMOD</em></td>
<td>Conference</td>
</tr>
<tr>
<td>$S_4$</td>
<td>2002</td>
<td>Year</td>
</tr>
</tbody>
</table>
Hand-Coded Methods

- Easy to construct in some cases
  - e.g., to recognize prices, phone numbers, zip codes, conference names, etc.
- Intuitive to debug and maintain
  - Especially if written in a “high-level” language:
    - Can incorporate domain knowledge
- Scalability issues:
  - Labor-intensive to create
  - Highly domain-specific
  - Often corpus-specific
  - Rule-matches can be expensive

ContactPattern ← RegularExpression(Email.body,”can be reached at”) [IBM Avatar]
Machine Learning Methods

- Can work well when lots of training data easy to construct

- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names
  - Non-local dependencies

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.™

[From AliBaba]
Any of these models can be used to capture words, formatting or both.
Popular Machine Learning Methods

For details: [Feldman, 2006 and Cohen, 2004]

- Naive Bayes
- SRV [Freitag 1998], Inductive Logic Programming
- Rapier [Califf and Mooney 1997]
- Hidden Markov Models [Leek 1997]
- Maximum Entropy Markov Models [McCallum et al. 2000]
- Conditional Random Fields [Lafferty et al. 2001]

- Scalability
  - Can be labor intensive to construct training data
  - At run time, complex features can be expensive to construct or process
    (batch algorithms can help: [Chandel et al. 2006] )
Some Available Entity Taggers

- **ABNER:**
  - http://www.cs.wisc.edu/~bsettles/abner/
  - Linear-chain conditional random fields (CRFs) with orthographic and contextual features.

- **Alias-I LingPipe**
  - http://www.alias-i.com/lingpipe/

- **Mallet:**
  - http://mallet.cs.umass.edu/index.php/Main_Page
  - Collection of NLP and ML tools, can be trained for name entity tagging

- **MinorThird:**
  - http://minorthird.sourceforge.net/
  - Tools for learning to extract entities, categorization, and some visualization

- **Stanford Named Entity Recognizer:**
  - CRF-based entity tagger with non-local features
• Statistical named entity tagger
  • Generative statistical model
    • Find most likely tags given lexical and linguistic features
    • Accuracy at (or near) state of the art on benchmark tasks

• Explicitly targets scalability:
  • ~100K tokens/second runtime on single PC
  • Pipelined extraction of entities
  • User-defined mentions, pronouns and stop list
    • Specified in a dictionary, left-to-right, longest match
  • Can be trained/bootstrapped on annotated corpora
Outline

- Overview of Information Extraction
  - Entity tagging
  - Relation extraction
  - Event extraction

- Scaling up Information Extraction
  - Focus on scaling up to large collections (where data mining and ML techniques shine)
  - Other dimensions of scalability
Relation Extraction Examples

- Extract tuples of entities that are related in predefined way

**Disease Outbreaks relation**

<table>
<thead>
<tr>
<th>Date</th>
<th>Disease Name</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan. 1995</td>
<td>Malaria</td>
<td>Ethiopia</td>
</tr>
<tr>
<td>July 1995</td>
<td>Mad Cow Disease</td>
<td>U.K.</td>
</tr>
<tr>
<td>Feb. 1995</td>
<td>Pneumonia</td>
<td>U.S.</td>
</tr>
</tbody>
</table>

*May 19 1995, Atlanta --* The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly *Ebola* epidemic in *Zaire*, is finding itself hard pressed to cope with the crisis.
Relation Extraction Approaches

Knowledge engineering

- Experts develop rules, patterns:
  - Can be defined over lexical items: “<company> located in <location>”
  - Over syntactic structures: “((Obj <company>) (Verb located) (*) (Subj <location>))”
- Sophisticated development/debugging environments:
  - Proteus, GATE

Machine learning

- Supervised: Train system over manually labeled data

- Partially-supervised: train system by bootstrapping from “seed” examples:
  - Agichtein & Gravano 2000, Etzioni et al., 2004, Yangarber & Grishman 2001, …
  - “Open” (no seeds): Sekine et al. 2006, Cafarella et al. 2007, Banko et al. 2007

- Hybrid or interactive systems:
  - Experts interact with machine learning algorithms (e.g., active learning family) to iteratively refine/extend rules and patterns
  - Interactions can involve annotating examples, modifying rules, or any combination
Traditional vs Open Information extraction

Table 2: The contrast between traditional and open IE.

<table>
<thead>
<tr>
<th></th>
<th>Traditional IE</th>
<th>Open IE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>Corpus + Labeled Data</td>
<td>Corpus + Domain-Independent Methods</td>
</tr>
<tr>
<td><strong>Relations</strong></td>
<td>Specified In Advance</td>
<td>Discovered Automatically</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>$O(D \times R)$</td>
<td>$O(D)$</td>
</tr>
<tr>
<td></td>
<td>$D$ documents, $R$ relations</td>
<td>$D$ documents</td>
</tr>
</tbody>
</table>
How open IE systems work

- learn a general model of how relations are expressed (in a particular language), based on unlexicalized features such as part-of-speech tags. (Identify a verb)

- Learn domain-independent regular expressions. (Punctuations, Commas).
Is there a general model of relationships in English

**Table 1: Taxonomy of binary relationships. Nearly 95% of 500 randomly selected sentences belong to one of the eight categories noted here.**

<table>
<thead>
<tr>
<th>Relative Frequency</th>
<th>Category</th>
<th>Simplified Lexico-Syntactic Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.8</td>
<td>Verb</td>
<td>(E_1) Verb (E_2) (X) established (Y)</td>
</tr>
<tr>
<td>22.8</td>
<td>Noun + Prep</td>
<td>(E_1) NP Prep (E_2) (X) settlement with (Y)</td>
</tr>
<tr>
<td>16.0</td>
<td>Verb + Prep</td>
<td>(E_1) Verb Prep (E_2) (X) moved to (Y)</td>
</tr>
<tr>
<td>9.4</td>
<td>Infinitive</td>
<td>(E_1) to Verb (E_2) (X) plans to acquire (Y)</td>
</tr>
<tr>
<td>5.2</td>
<td>Modifier</td>
<td>(E_1) Verb (E_2) Noun (X) is (Y) winner</td>
</tr>
<tr>
<td>1.8</td>
<td>Coordinate_n</td>
<td>(E_1) (and,</td>
</tr>
<tr>
<td>1.0</td>
<td>Coordinate_y</td>
<td>(E_1) (and,</td>
</tr>
<tr>
<td>0.8</td>
<td>Appositive</td>
<td>(E_1) NP ((\cdot)) (Y) (X) hometown : (Y)</td>
</tr>
</tbody>
</table>
The Self-Annotations Web

- There is a huge amount of implicit knowledge in the Web
- Make use of this implicit knowledge together with statistical information to propose formal annotations and overcome the vicious cycle:
  $$\text{semantics} \approx \text{syntax} + \text{statistics}?$$
- Annotation by maximal statistical evidence
A small quiz

What is Laksa?

A: dish    B: city

C: temple  D: mountain
Asking Google!

- „cities such as Laksa“ 0 hits
- „dishes such as Laksa“ 10 hits
- „mountains such as Laksa“ 0 hits
- „temples such as Laksa“ 0 hits

⇒ Google knows more than all of you together!
⇒ Example of using syntactic information + statistics to derive semantic information
Patterns

- HEARST1: <CONCEPT>s such as <INSTANCE>
- HEARST2: such <CONCEPT>s as <INSTANCE>
- HEARST3: <CONCEPT>s, (especially/including) <INSTANCE>
- HEARST4: <INSTANCE> (and/or) other <CONCEPT>s

Examples:
- dishes such as Laksa
- such dishes as Laksa
- dishes, especially Laksa
- dishes, including Laksa
- Laksa and other dishes
- Laksa or other dishes
Patterns (Cont’d)

- DEFINITE1: the <INSTANCE> <CONCEPT>
- DEFINITE2: the <CONCEPT> <INSTANCE>

- APPOSITION: <INSTANCE>, a <CONCEPT>
- COPULA: <INSTANCE> is a <CONCEPT>

Examples:
- the Laksa dish
- the dish Laksa
- Laksa, a dish
- Laksa is a dish
Asking Google (more formally)

- Instance $i \in I$, concept $c \in C$, pattern $p \in \{\text{Hearst1}, \ldots, \text{Copula}\}$ $\text{count}(i,c,p)$ returns the number of Google hits of instantiated pattern

$$\text{count}(i,c) := \sum_p \text{count}(i,c,p)$$

- E.g. $\text{count}(\text{Laksa}, \text{dish}) := \text{count}(\text{Laksa}, \text{dish}, \text{def1}) + \ldots$

- Restrict to the best ones beyond threshold $\theta$

$$R_\theta := \left\{(i,c_i) \mid i \in I, c_i := \arg \max_{c \in C} \text{count}(i,c) \land \text{count}(i,c) \geq \theta \right\}$$
<table>
<thead>
<tr>
<th>Example</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic city</td>
<td>1520837</td>
</tr>
<tr>
<td>Bahamas island</td>
<td>649166</td>
</tr>
<tr>
<td>USA country</td>
<td>582275</td>
</tr>
<tr>
<td>Connecticut state</td>
<td>302814</td>
</tr>
<tr>
<td>Caribbean sea</td>
<td>227279</td>
</tr>
<tr>
<td>Mediterranean sea</td>
<td>212284</td>
</tr>
<tr>
<td>Canada country</td>
<td>176783</td>
</tr>
<tr>
<td>Guatemala city</td>
<td>174439</td>
</tr>
<tr>
<td>Africa region</td>
<td>131063</td>
</tr>
<tr>
<td>Australia country</td>
<td>128607</td>
</tr>
<tr>
<td>France country</td>
<td>125863</td>
</tr>
<tr>
<td>Germany country</td>
<td>124421</td>
</tr>
<tr>
<td>Easter island</td>
<td>96585</td>
</tr>
<tr>
<td>St Lawrence river</td>
<td>65095</td>
</tr>
<tr>
<td>Commonwealth state</td>
<td>49692</td>
</tr>
<tr>
<td>New Zealand island</td>
<td>40711</td>
</tr>
<tr>
<td>Adriatic sea</td>
<td>39726</td>
</tr>
<tr>
<td>Netherlands country</td>
<td>37926</td>
</tr>
<tr>
<td>St John church</td>
<td>34021</td>
</tr>
<tr>
<td>Belgium country</td>
<td>33847</td>
</tr>
<tr>
<td>San Juan island</td>
<td>31994</td>
</tr>
<tr>
<td>Mayotte island</td>
<td>31540</td>
</tr>
<tr>
<td>EU country</td>
<td>28035</td>
</tr>
<tr>
<td>UNESCO organization</td>
<td>27739</td>
</tr>
<tr>
<td>Austria group</td>
<td>24266</td>
</tr>
<tr>
<td>Greece island</td>
<td>23021</td>
</tr>
<tr>
<td>Malawi lake</td>
<td>21081</td>
</tr>
<tr>
<td>Israel country</td>
<td>19732</td>
</tr>
<tr>
<td>Perth street</td>
<td>17880</td>
</tr>
<tr>
<td>Luxembourg city</td>
<td>16393</td>
</tr>
<tr>
<td>Nigeria state</td>
<td>15650</td>
</tr>
<tr>
<td>St Croix river</td>
<td>14952</td>
</tr>
<tr>
<td>Nakuru lake</td>
<td>14840</td>
</tr>
<tr>
<td>Kenya country</td>
<td>14382</td>
</tr>
<tr>
<td>Benin city</td>
<td>14126</td>
</tr>
<tr>
<td>Cape Town city</td>
<td>13768</td>
</tr>
</tbody>
</table>
Evaluation Scenario

- Corpus: 45 texts from http://www.lonelyplanet.com/destinations
- Ontology: tourism ontology from GETESS project
  - #concepts: original – 1043; pruned – 682
- Manual Annotation by two subjects:
  - A: 436 instance/concept assignments
  - B: 392 instance/concept assignments
  - Overlap: 277 instances (Gold Standard)
  - A and B used 59 different concepts
  - Categorial (Kappa) agreement on 277 instances:
    \( \kappa = 63.5\% \)
Cohen’s Kappa

- **Cohens Kappa** for interrater reliability

\[
\kappa = \frac{p_{agree} - p_{random}}{1 - p_{random}}
\]

- \( p_{agree} \): which fraction of times to raters agree
- \( p_{random} \): which fraction of times would they agree by chance

- \( \kappa = 0 \) \( \Rightarrow \) no agreement
- \( \kappa = 1 \) \( \Rightarrow \) perfect agreement
- \( \kappa < 0 \) \( \Rightarrow \) tending towards disagreement
Precision and Recall

- Correct decisions:
  1. correct alarms (ca)
  2. correct dismissals (cd)

- Erroneous decisions:
  1. false alarms (fa)
  2. false dismissals (fd)

- \(fa, fd, ca, cd\) represent the numbers of times that documents from a result set fall into any of the four possible categories
Item set

<table>
<thead>
<tr>
<th>User judgment</th>
<th>System judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>relevant</td>
<td>relevant, irrelevant</td>
</tr>
<tr>
<td>irrelevant</td>
<td>cd, fd</td>
</tr>
</tbody>
</table>

- cd
- fd
- ca
- fa
**Precision / Recall**

Item set

\[
\begin{align*}
    \text{Prec} &= \frac{ca}{ca + fa} \in [0, 1] \\
    \text{Rec} &= \frac{ca}{ca + fd} \in [0, 1]
\end{align*}
\]
### Example

<table>
<thead>
<tr>
<th></th>
<th>User 1</th>
<th>User 2</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Niger: country</td>
<td>Niger: river</td>
<td>Niger: country</td>
</tr>
<tr>
<td>2</td>
<td>Germany: country</td>
<td>Germany: country</td>
<td>Germany: country</td>
</tr>
<tr>
<td>3</td>
<td>Main: river</td>
<td>Main: river</td>
<td>Main: street</td>
</tr>
<tr>
<td>4</td>
<td>India: country</td>
<td>India: region</td>
<td>India: country</td>
</tr>
</tbody>
</table>

\[ \kappa = \left( \frac{1}{2} - \frac{1}{3} \right) / 1 - \left( \frac{1}{3} \right) = \frac{1}{4} \]

Assume Gold Standard: User 1

Then: \( ca = 3 \)
      \( fd = 1 \)
      \( fa = 1 \)

\( \text{Prec} = \frac{3}{3+1} = \frac{3}{4} \)
\( \text{Rec} = \frac{3}{3+1} = \frac{3}{4} \)

Assume Gold Standard: User 2

Then: \( ca = 1 \)
      \( fd = 3 \)
      \( fa = 3 \)

Open world: often Prec@10 and Recall unknown
Results

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2824</td>
<td>0.2490</td>
<td>0.2824</td>
</tr>
</tbody>
</table>

F = 28.24%  
R/Acc = 24.90%
PANKOW & CREAM/OntoMat

ISWeb - Information Systems & Semantic Web
Maciej Janik
janik@uni-koblenz.de
Semantic Web
40 of 66
Information Extraction – Extracting Knowledge from Wikipedia

Steffen Staab
Maciej Janik

Semantic Web
2009-07-10
Extracting knowledge

- Sören Auer, Jens Lehmann. „What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content.” ESWC 2007

- The amount of knowledge in Wikipedia
  - English ~ 3M articles
  - German ~1M articles
  - Polish ~620K articles
  - Over 250 languages
  - Created and edited by users, for other users to read
  - Arranged into categories, included some structure
  - … but generally not machine understandable
- DBpedia is a community effort to
  - extract structured ("infobox") information from Wikipedia
  - provide a query endpoint to the dataset
  - interlink the DBpedia dataset with other datasets on the Web

- Why it works?
  - Data: Wikipedia full dumps freely available
  - Use of open source software
Wikipedia Templates

- Structured and “standardized” knowledge in Wikipedia
- Reused for multiple objects that share similar semantics

```wikipedia
{{Infobox Town AT |
name = Innsbruck |
image_coa = InnsbruckWappen.png |
image_map = Karte-tirol-1.png |
state = [[Tyrol]] |
regbzk = [[Statutory city]] |
population = 117,342 |
population_as_of = 2006 |
pop_dens = 1,119 |
area = 104.91 |
elevation = 574 |
lat_deg = 47 |
lat_min = 16 |
lat_hem = N |
lon_deg = 11 |
lon_min = 23 |
lon_hem = E |
p postal_code = 6010-6080 |
area_code = 0512 |
licence = I |
mayor = Hilde Zach |
website = [http://innsbruck.at] |
}}
```
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpprop:area</td>
<td>104.91 (xsd:double)</td>
</tr>
<tr>
<td>dbpprop:areaCode</td>
<td>512 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:district</td>
<td>Statutory city (en)</td>
</tr>
<tr>
<td>dbpprop:elevation</td>
<td>574 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:hasPhotoCollection</td>
<td><a href="http://www4.wiwiss.fu-berlin.de/flickrwrap/photos/Innsbruck">http://www4.wiwiss.fu-berlin.de/flickrwrap/photos/Innsbruck</a></td>
</tr>
<tr>
<td>dbpprop:imageCaption</td>
<td>Innsbruck (en)</td>
</tr>
<tr>
<td>dbpprop:imagePhoto</td>
<td><a href="http://upload.wikimedia.org/wikipedia/commons/0/07/IMG_9039-Innsbruck.JPG">http://upload.wikimedia.org/wikipedia/commons/0/07/IMG_9039-Innsbruck.JPG</a></td>
</tr>
<tr>
<td>dbpprop:imagesize</td>
<td>300px (en)</td>
</tr>
<tr>
<td>dbpprop:latDeg</td>
<td>47 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:latHem</td>
<td>N (en)</td>
</tr>
<tr>
<td>dbpprop:latMin</td>
<td>16 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:licence</td>
<td>I (en)</td>
</tr>
<tr>
<td>dbpprop:lonDeg</td>
<td>11 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:lonHem</td>
<td>E (en)</td>
</tr>
<tr>
<td>dbpprop:lonMin</td>
<td>23 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:mayor</td>
<td>dbpedia:Hilde_Zach</td>
</tr>
<tr>
<td>dbpprop:name</td>
<td>Innsbruck (en)</td>
</tr>
<tr>
<td>dbpprop:popDens</td>
<td>1119 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:population</td>
<td>117916 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:populationAsOf</td>
<td>01.01.2007 (en)</td>
</tr>
<tr>
<td>dbpprop:postalCode</td>
<td>6010-6080 (en)</td>
</tr>
</tbody>
</table>
Wiki Templates in extraction – algorithm

- Identify pages with templates

- Choose well-populated / used templates
  - Unused templates can only bring errors

- Parse template (XMLized Wikipedia format)
  - Get attributes → relations
  - Get values → objects

- Create relevant triples from extracted information
  - URI-fy references
  - Add data types (if known from template context)
Templates – Obstacles in extraction

- Not all pages include templates
  - Even within the same topic not all pages may have it

- Template definition flaws
  - Not well-formed – include presentation properties
  - Multiple values as objects for more intuitive presentation
    - [[Innsbruck]], [[Austria]]
  - Complex and redundant attribute values
    - height=5’11” (180cm)
  - One subject can have multiple templates defined
    - Infobox_Film, Infobox Film, Infobox_film, ...
  - Multiple attribute names to define the same relationship
Further facts

- Exploiting rich Wikipedia linking
  - Unnamed links to other entities in Wikipedia
    - HREFs (for now …)
  - Categorization
    - Wikipedia categories
    - Not strict hierarchy, rather thesaurus
  - Multiple languages
    - Variety of languages for the same entity
  - Links to open data and other web resources
    - e.g. geo locations
  - Links to other Wiki projects
    - Wiktionary
    - Wikimedia Commons
Extracting structured data from Wikipedia

Maciej Janik  
janik@uni-koblenz.de

Semantic Web  
49 of 66
Automatic links among open datasets

Processors can switch automatically from one to the other…
Extracted Wikipedia
→ DBpedia.org
Linked Open Data
SPARQL Query Interface
Johan August Strindberg (pronounced [ɛŋˈwɪkstːɛːn] (1849-1912) was a Swedish playwright and writer. He is arguably the most influential and important of all Swedish authors, and one of the most influential Scandinavian authors, along with Knut Hamsun, with whom he fraternized while in Paris in the mid 1890s. Henrik Ibsen, Søren Kierkegaard and Hans Christian Andersen. Strindberg is known as one of the fathers of modern theatre. His work falls into two major literary movements, Naturalism and Expressionism.[1]

This article needs additional citations for verification. Please help improve this article by adding reliable references. Unsourced material may be challenged and removed. (December 2006)

**Biography**

**Early years**

Strindberg was the third son of Carl Oscar Strindberg, a shipping agent, and Ulrika Eleonora (Nordiq, Noring). Ulrika was twelve years Carl's junior and of humble origin, called a "domestic servant woman" by Strindberg. He used this
1. Use URIs as names for things
2. Use http URIs
3. When someone looks up a name, provide useful information
4. Include links to other URIs so that they can discover more things
<table>
<thead>
<tr>
<th>property</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (en)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (fr)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>ヨハン・アウグスト・ストリンデンベルリ</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (nl)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (pl)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>Стринберг, Юхан Август (ru)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (sv)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (fi)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (it)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (pt)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>奥古斯特 斯特林堡 (zh)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (da)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (de)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>August Strindberg (no)</td>
</tr>
<tr>
<td>owl:sameAs</td>
<td><a href="http://www4.wiwiss.fu-berlin.de/gutendata/resource/people/Strindberg">http://www4.wiwiss.fu-berlin.de/gutendata/resource/people/Strindberg</a> August_1849-1</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:1912_deaths</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:1849_births</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Alchemists</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Expressionist_dramatists_and_playwrights</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:French-language_writers</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Modernist_drama%2C_theatre_and_performance</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:People_from_Stockholm</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Swedish_language_writers</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Swedish_dramatists_and_playwrights</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Swedish_socialists</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Swedish_writers</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Swedish_novelists</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Uppsala_University_alumni</td>
</tr>
<tr>
<td>skos:subject</td>
<td>dbpedia:Category:Naturalist_dramatists_and_playwrights</td>
</tr>
<tr>
<td>foaf:depiction</td>
<td><a href="http://upload.wikimedia.org/wikipedia/commons/thumb/4/49/August_Strindberg.jpg/210px-August_Strindberg.jpg">http://upload.wikimedia.org/wikipedia/commons/thumb/4/49/August_Strindberg.jpg/210px-August_Strindberg.jpg</a></td>
</tr>
<tr>
<td>foaf:img</td>
<td><a href="http://upload.wikimedia.org/wikipedia/commons/4/49/August_Strindberg.jpg">http://upload.wikimedia.org/wikipedia/commons/4/49/August_Strindberg.jpg</a></td>
</tr>
<tr>
<td>foaf:name</td>
<td>August Strindberg</td>
</tr>
<tr>
<td>foaf:page</td>
<td><a href="http://en.wikipedia.org/wiki/August_Strindberg">http://en.wikipedia.org/wiki/August_Strindberg</a></td>
</tr>
<tr>
<td>is dbpedia-owl:author of</td>
<td>dbpedia:Inferno_%28Strindberg%29</td>
</tr>
<tr>
<td>is dbpedia-owl:influenced of</td>
<td>dbpedia:Emanuel_Swedenborg</td>
</tr>
<tr>
<td>is dbpedia-owl:influenced of</td>
<td>dbpedia:Otto_Weininger</td>
</tr>
<tr>
<td>is dbpedia-owl:influences of</td>
<td>dbpedia:S%C3%B6ren_Kierkegaard</td>
</tr>
<tr>
<td>is dbpedia-owl:influences of</td>
<td>dbpedia:Tennessee_Williams</td>
</tr>
<tr>
<td>is dbpedia-owl:writer of</td>
<td>dbpedia:Strindberg%2C_theatre_and_performance</td>
</tr>
<tr>
<td>is dbpedia-owl:author of</td>
<td>dbpedia:Strindberg%2C_theatre_and_performance</td>
</tr>
</tbody>
</table>

Include links to other URIs so that they can discover more things.

cont. http://dbpedia.org/page/August_Strindberg
Johan August Strindberg was a Swedish writer, playwright, and painter. Along with Knut Hamsun, with whom he frequented Ibsen, Søren Kierkegaard and Hans Christian Andersen he is arguably the most influential and most important of all the fathers of modern theatre. His work falls into two major literary movements, Naturalism and Expressionism.

About: August Strindberg
An Entity in Data Space: dbpedia.org

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:birthdate</td>
<td>1849-01-21 (xsd:date)</td>
</tr>
<tr>
<td>dbpedia-owl:birthplace</td>
<td>dbpedia:Stockholm</td>
</tr>
<tr>
<td>dbpedia-owl:deathplace</td>
<td>dbpedia:Stockholm</td>
</tr>
<tr>
<td>dbpedia-owl:deathdate</td>
<td>1912-05-14 (xsd:date)</td>
</tr>
<tr>
<td>dbpedia-owl:movement</td>
<td>dbpedia:Expressionism, dbpedia:Naturalism, dbpedia:Painting</td>
</tr>
<tr>
<td>dbpedia-owl:occupation</td>
<td>dbpedia:Playwright, dbpedia:Writer</td>
</tr>
</tbody>
</table>

http://dbpedia.org/page/August_Strindberg
Stockholm is Sweden’s capital and its largest city. It is the site of the national Swedish government, the parliament, and the official residence of the Swedish monarch. As of 2003, it is home to 21.3% of the Swedish population and contributes 29.1% of Sweden’s gross domestic product.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:establishedTitle</td>
<td>First mention</td>
</tr>
<tr>
<td>dbpedia-owl:foundingDate</td>
<td>1262</td>
</tr>
<tr>
<td>dbpedia-owl:latitudeminutes</td>
<td>21</td>
</tr>
<tr>
<td>dbpedia-owl:latitudenorthsouth</td>
<td>N</td>
</tr>
<tr>
<td>dbpedia-owl:latitudedegrees</td>
<td>59</td>
</tr>
<tr>
<td>dbpedia-owl:leaderName</td>
<td>dbpedia:Moderate_Party</td>
</tr>
<tr>
<td></td>
<td>dbpedia:Stan_Nordén</td>
</tr>
<tr>
<td>dbpedia-owl:leaderTitle</td>
<td>Mayor</td>
</tr>
<tr>
<td>dbpedia-owl:longitudedegrees</td>
<td>18</td>
</tr>
<tr>
<td>dbpedia-owl:longitudeminutes</td>
<td>4</td>
</tr>
<tr>
<td>dbpedia-owl:longitudeseastwest</td>
<td>E</td>
</tr>
<tr>
<td>dbpedia-owl:country</td>
<td>Stockholm</td>
</tr>
<tr>
<td>dbpedia-owl:countryEn</td>
<td>Stad</td>
</tr>
<tr>
<td>dbpedia-owl:countryEn</td>
<td>Stockholm</td>
</tr>
<tr>
<td>dbpedia-owl:populationAsOf</td>
<td>2008</td>
</tr>
<tr>
<td>dbpedia-owl:populationMetro</td>
<td>1949515</td>
</tr>
<tr>
<td>dbpedia-owl:populationTotal</td>
<td>802611</td>
</tr>
<tr>
<td>dbpedia-owl:populationUrbanTotal</td>
<td>1252020</td>
</tr>
<tr>
<td>p:abstract</td>
<td>Stockholm is Sweden’s capital and its largest city. It is the site of the national Swedish government, the parliament, and the official residence of the Swedish monarch. As of 2003, it is home to 21.3% of the Swedish population and contributes 29.1% of Sweden’s gross domestic product.</td>
</tr>
</tbody>
</table>
Extracting additional facts

- Templates cover only very small fraction of knowledge

- Named links in Wikipedia do not provide all required relationships

- Large knowledge resides within “plain” HREFs
  … but this knowledge is hidden in text created for humans

Need to parse free text to get meaningful relationships
More Patterns…. 
- Rule and Heuristic based method
  - Pattern-based approach
  - Uses WordNet

- YAGO [Suchanek et. al, 2007]
- [Ruiz-Casado et. al. 2006]
- [Weld et al]
- [Suchanek et al 2009]
Extracting more semantics

- Learning common patterns for interesting relationships
  - is-author-of, is-the-capital-of, is-employee-of, ...

- Seed initial patterns and look for them in training corpus
- Extract common patterns between linked entries
  - Search for support in Wikipedia
  - Search for pattern support using search engines
- Generalize it
- Apply it to extract more information
Applying open information extraction on Wikipedia

List of related word pairs → List of patterns → Generalisation, pruning and disambiguation → Generalised and pruned pattern set

Web download → Extraction of sentences

Training corpus

Set of new extracted pairs that hold the relation

Extraction of related pairs
Generalizing patterns

- Example (known entities are underlined):
  - **Alfred Hitchcock** directed the famous film *Psycho*
  - **Alfred Hitchcock** directed the well known film *Psycho*

  - ?x directed the famous film ?y
  - ?x directed the well known film ?y

  - ?x directed the famous | well known film ?y

  - ?x directed the * famous | known film ?y

- Apply pattern
  - **Alfred Hitchcock** directed the famous film *The Birds*
  - **Bernardo Bertolucci** directed the well known film *The Last Emperor*
  - **Woody Allen** directed the amusing and famous film *Annie Hall*
Finding good patterns

- Good patterns
  - Support in training data
  - Found in Wikipedia documents
  - Unambiguous – matching same types / topics
    - Not the best one: \(?x’s \ ?y
      - Einstein’s Theory of General Relativity
      - Bosco’s The Garden of Delights
      - Tolkien’s Lord of the Rings
    - Need pruning
  - Support in free text search
    - Check how often the pairs you found are matched with this pattern in web documents
WordNet

- Lexical database for the English language
- Created at the Cognitive Science Laboratory of Princeton University
- Groups English words into sets of synonyms called synsets
- Provides short, general definitions
- Provides hypernym/hyponym relations
  - e.g. canine is hypernym, dog is hyponym
WordNet Search - 3.0

Word to search for: dog
Display Options: (Select option to change) Change

Key: "S." = Show Synset (semantic) relations, "W." = Show Word (lexical) relations

Noun

(42) S: (n) dog, domestic dog, Canis familiaris
   ○ direct hypernym / full hypernym
       S: (n) puppy
       S: (n) pooch, doggie, doggy, barker, bow-wow
       S: (n) cur, mongrel, mutt
       S: (n) lap dog
   ○ direct hypernym / inherited hypernym / sister term
       S: (n) canine, canid
       S: (n) domestic animal, domesticated animal
Type Extraction with WordNet

- Goal: create class hierarchy
  - e.g. singer subClassOf performer
    performer subClassOf artist
- hyponymy relation from WordNet
- Wikipedia class ‘American people in Japan’ is subclass of WordNet class ‘person’