CS544: Semantic Class Learning

February 14, 2012

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

How are Max Planck, Angela Merkel and Dalai Lama related?

All have doctoral degrees from German universities
The problem with automated Question Answering

- Where do lobsters like to live?
  — on the table

- Where are zebras most likely found?
  — in the dictionary

- What is an invertebrate?
  — Dukakis

Michael Dukakis is a member of the Democratic Party, I have long suspected that elected officials from the Democratic Party are some previously unclassified form of invertebrate, a totally spineless creature capable of great noise but no real movement or action

Solve our toy example & help your NER homework

- researcher
- teacher
- Semantic Class Learning
- killer
- lawyer

Improve Accuracy
Semantic Class

- **Definition:** A semantic class contains words that share a semantic property.

- **Example:**
  - *Nobel Prize Winners:* \{*Albert Einstein, Max Plank ...\}
  - *Russian Federation States:* \{*Georgia, Ukrain, Russia ...\}
  - **People:** \{*teacher, student, girl, boy, Will Smith ...\}

---

### Ontologies

```
entity
   subclassOf
    person
        subclassOf
         scientists
         subclassOf
          singer
             type
              "Elvis"

   subclassOf
    location
        subclassOf
         city
            type
              "Tupelo"

   subclassOf
    person
        label
         "The King"

The same label for two entities: *homonymy*

The same entity has two labels: *synonymy*
```

---

**Classes**

**Relations**

**Instances**

**Labels/words**
Characteristics of Semantic Classes

- Based on instance membership
  - closed (i.e. fixed size membership)
    - small (names of countries, states, planets)
    - large (names of diseases, cities)
  - open (i.e. the size of the members constantly varies)
    Example. singers, movie titles

- An instance can belong to multiple classes
  Ex. orange can be a fruit and a colour

- Instances and classes are bound by diverse relations
Where do Semantic Classes, Instances and Relations reside?
WordNet

What if we could make the English language computer-processable?

George Miller

- started in 1985
- Cognitive Science Laboratory, Princeton University
- written by lexicographers
- goal: support automatic text analysis and AI applications

[Miller, CACM 1995]
WordNet: Semantic Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy (N, V, Adj, Adv)</td>
<td>Same sense</td>
<td>(camera, photographic camera)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(fast, speedy)</td>
</tr>
<tr>
<td>Antonymy (Adj, Adv)</td>
<td>Opposite</td>
<td>(fast, slow)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(buy, sell)</td>
</tr>
<tr>
<td>Hypernymy (N)</td>
<td>Is-A</td>
<td>(rabbit, animal)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(gun, weapon)</td>
</tr>
<tr>
<td>Meronymy (N)</td>
<td>Part</td>
<td>(camera, optical lens)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(car, suspension)</td>
</tr>
<tr>
<td>Troponymy (V)</td>
<td>Manner</td>
<td>(buy, subscribe)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(sell, retail)</td>
</tr>
<tr>
<td>Entailment (V)</td>
<td>X must mean doing Y</td>
<td>(buy, pay)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(sell, give)</td>
</tr>
</tbody>
</table>
# Exercise

- Fill in the semantic relations for **DOG**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy (N, V, Adj, Adv)</td>
<td>Same sense</td>
<td></td>
</tr>
<tr>
<td>Antonymy (Adj, Adv)</td>
<td>Opposite</td>
<td></td>
</tr>
<tr>
<td>Hyponymy (N)</td>
<td>Is-A</td>
<td></td>
</tr>
<tr>
<td>Meronymy (N)</td>
<td>Part</td>
<td></td>
</tr>
<tr>
<td>Troponymy (V)</td>
<td>Manner</td>
<td></td>
</tr>
<tr>
<td>Entailment (V)</td>
<td>X must mean doing Y</td>
<td></td>
</tr>
</tbody>
</table>
WordNet: Is-A Relation Hierarchy

animal
chordate
vertebrate
mammal
placental
carnivore
canine
dog

Do you notice any problems or limitations with WordNet?

But WordNet is not enough ...

• Has limited coverage
  – many instances and classes are missing
  – not all relations are listed
  – knowledge does not cover all domains
  – hierarchy representation
Example of WordNet’s Limitations

• If you want to know:
  – *all names of US presidents, you will notice that the name of Barack Obama is not present*

  – *the names of the Chinese, Indian, French presidents, you will notice that neither these classes nor their instances are present at all*

  – *more information on Boo the dog, you have higher chances of finding them on Facebook than WordNet*

• WordNet has more information on **animals** than **people**
So how can we fix WordNet?

Let’s start with a realistic example

• Build an automated system that can learn all:
  – US states
  – country names
  – and cities in the world
Automatic Semantic Class Learning

- Task Definition: given a class and an instance, learn automatically with minimum supervision
  - instances
  - classes
  - ISA relations
  - taxonomic organization

How would you initiate the learning process?

Lexico-Syntactic Patterns (Hearst 92)

(S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.
Lexico-Syntactic Patterns (Hearst 92)

(S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

(1a) NP₀ such as NP₁ {, NP₂ ... , (and | or) NPᵢ} i ≥ 1 are such that they imply

(1b) for all NPᵢ, i ≥ 1, hyponym(NPᵢ, NP₀)

Thus from sentence (S1) we conclude

hyponym(“Gelidium”, “red algae”).

Examples are adapted from Marti Hearst
Lexico-Syntactic Patterns (Hearst 92)

(2) such NP as \{NP ,\} * {(or | and)} NP

... works by such authors as Herrick, Goldsmith, and Shakespeare.

⇒ hyponym(“author”, “Herrick”),

hyponym(“author”, “Goldsmith”),

hyponym(“author”, “Shakespeare”)

Lexico-Syntactic Patterns (Hearst 92)

(2) such NP as \{NP ,\} * {(or | and)} NP

... works by such authors as Herrick, Goldsmith, and Shakespeare.

⇒ hyponym(“author”, “Herrick”),

hyponym(“author”, “Goldsmith”),

hyponym(“author”, “Shakespeare”)

(3) NP \{ , NP \} * {,} or other NP

Bruises, ..., broken bones or other injuries ...

⇒ hyponym(“bruise”, “injury”),

hyponym(“broken bone”, “injury”)

Examples are adapted from Marti Hearst
Examples

• **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**

• I listen to pretty much all music but prefer **country** such as **Garth Brooks**

Lesson Learned

• To acquire knowledge, we need patterns
• ... but not all extractions and patterns are reliable

• What makes a pattern extraction to be good?
  – extractions should **occur frequently in text**
  – extractions should (nearly) always suggest the relation of interest
  – extractions should be recognizable with little pre-encoded knowledge
KnowItAll Architecture (Etzioni et al.05)

- use many patterns
- use synonyms of class terms
- rank instances

Lets Learn City Names

- Input:
  - search query:
    - “city; town”, “cities; towns”
  - use all extraction rules following (Hearst 92):
    - <class2> such as <NPList>
    - <NP> is a <class1>
    - <class2> including <NPList>

- Generate extraction queries for any search engine:
  - “cities such as”
  - “is a town”
  - “towns including”
Learning City Names

- Submit extraction queries to Google and collect the returned snippets:

- **Central Highlands Council - Welcome - Enjoy the historic buildings ...**
  Enjoy historic buildings and friendly towns including Bothwell, Hamilton, Gretna and Ellendale to name a few. Fish at great fishing spots.
  www.centralhighlands.tas.gov.au - Cached - Similar

- **Wichita, Kansas RE/MAX Agent serving Wichita and surrounding towns ...**
  Wichita, Kansas RE/MAX realtor serving Wichita, Goddard, Maize, Bentley, Halstead, Sedgwick, Park City, Valley Center, Bel Aire, Andover, Derby, Rose Hill, ...
  www.wichitarealestate4you.net - Cached

- **Public Health And Poor-Law Medical Services**
  towns, including London. 6,344 births and 5,067 deaths were registered during the week ending Saturday, July 25th. The annual rate of mortality ...
  www.jstor.org/stable/2036873

- **John D. Williams, M.D., B.Sc.Edin., Honorary Gynaecologist To The ...**
  by JW B - 1901
towns, including London. 6561 births and 3674 deaths were registered during the week ending Saturday last, May 25th. The annual rate of mortality ...
  www.jstor.org/stable/20268562

- **Sanitary and meteorological notes**
  annually of 21°2 in twenty-eight large English towns (including London, in which the rate was 19°7), 30°8 in the sixteen chief towns of Ireland, ...
  www.springerlink.com/index/30401P77H45848X.pdf

Extracting City Names

- Pull all *candidate* city names from the snippets using extraction rules

**Central Highlands Council - Welcome - Enjoy the historic buildings ...**
Enjoy historic buildings and friendly towns including Bothwell, Hamilton, Gretna and Ellendale to name a few. Fish at great fishing spots.

```xml
<class2> including <NPList>
Bothwell
Hamilton
Gretna
Ellendale
```
Assessing Candidates

• **STEP1**: Generate *discriminators* from the rules and the user input
  – cities such as `<Candidate>`
  – `<Candidate>` is a town
  – `<Candidate>` is a city
  – towns including `<Candidate>`

Assessing Candidates

• **STEP2**: Generate *discriminator queries* from the discriminators and the extracted candidates
  – cities such as *London*
  – *London* is a town
  – *London* is a city
  – towns including *London*
Assessing Candidates

- **STEP3**: Evaluate each candidate instance with each discriminator query and compute PMI as:

\[
PMI(Cnd, Disc) = \frac{|\text{Hits}(\text{DiscriminatorTerm} + \text{CandidateInstance})|}{|\text{Hits}(\text{CandidateInstance})|}
\]

\[
PMI(\text{London, city}) = \frac{\text{Hits(\text{city London})}}{\text{Hits(\text{London})}} = \frac{8,590,000}{533,000,000} = 0.0161
\]

\[
PMI(\text{Avocado, city}) = \frac{\text{Hits(\text{city Avocado})}}{\text{Hits(\text{Avocado})}} = \frac{5,980}{8,320,000} = 0.000718
\]

\[PMI(\text{London, city}) >> PMI(\text{Avocado, city})\]
Learning Class Names

- Use the same patterns in inverse order to acquire class names for a given instance

![Diagram](image)

- How would we estimate the goodness of a class name?

\[ \chi^2 \text{ statistic (CHI)} \]

Is “jaguar” a good predictor for the “auto” class?

<table>
<thead>
<tr>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = auto</td>
<td>2</td>
</tr>
<tr>
<td>Class ≠ auto</td>
<td>3</td>
</tr>
</tbody>
</table>

We want to compare:
- the observed distribution above; and
- null hypothesis: that jaguar and auto are independent
\[ \chi^2 \text{ statistic (CHI)} \]

- Null hypothesis says “jaguar” and “auto” are independent
- How many co-occurrences of “jaguar” and “auto” do we expect?
  - If independent: \( \text{Pr}(\text{jaguar,auto}) = \text{Pr}(\text{jaguar}) \times \text{Pr}(\text{auto}) \)
  - So, there would be: \( N \times \text{Pr}(\text{jaguar,auto}), \) i.e. \( N \times \text{Pr}(\text{jaguar}) \times \text{Pr}(\text{auto}) \)
    \[ \text{Pr}(\text{jaguar}) = \frac{2+3}{N}; \]
    \[ \text{Pr}(\text{auto}) = \frac{2+500}{N}; \]
    \[ N = 2 + 3 + 500 + 9500 \]
  - Which = \( N \times \frac{5}{N} \times \frac{502}{N} = 2510/N = 2510/10005 \approx 0.25 \)

<table>
<thead>
<tr>
<th></th>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = auto</td>
<td>2</td>
<td>500</td>
</tr>
<tr>
<td>Class ≠ auto</td>
<td>3</td>
<td>9500</td>
</tr>
</tbody>
</table>
\( \chi^2 \) statistic (CHI)

\( \chi^2 \) is interested in \((f_0 - f_e)^2/f_e\) summed over all table entries:

\[
\chi^2(jaguar, auto) = \sum (O - E)^2 / E
\]

<table>
<thead>
<tr>
<th>Class</th>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = auto</td>
<td>2 (0.25)</td>
<td>500 (502)</td>
</tr>
<tr>
<td>Class ≠ auto</td>
<td>3 (4.75)</td>
<td>9500 (9498)</td>
</tr>
</tbody>
</table>

The null hypothesis is rejected with confidence .999 since 12.9 > 10.83 (the value for .999 confidence).
$\chi^2$ statistic (CHI)

• There is a simpler formula for $\chi^2$:

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>#$(t, c)$</td>
</tr>
<tr>
<td>C</td>
<td>#$(\neg t, c)$</td>
</tr>
<tr>
<td>B</td>
<td>#$(t, \neg c)$</td>
</tr>
<tr>
<td>D</td>
<td>#$(\neg t, \neg c)$</td>
</tr>
</tbody>
</table>

$N = A + B + C + D$

Limitations

• For each query, the Search Engines return maximum 1000 snippets

• Pointwise Mutual Information and Chi-statistics are great but not always very sensitive for ranking
PROBLEM SOLVING SESSION I

Topics

• Language Identification
• Authorship Identification
• POS-tagging