Using WordNet glosses to refine Google queries

Jan Nemrava

Department of Information and Knowledge Engineering,
University of Economics, Prague
nemrava@vse.cz
The Present

- Web search results affected by synonymy
- Inexperienced users inserting too general queries
- Ambiguity of inserted queries
- Goal: Discover synonymy in proper nouns
- Pankow inspiration
Some of solutions

- Offer query refinement
  - Zoom in
  - Zoom out
    - Ask Jeeves (ask.com)

- Term suggestions
  - Not grammatically and meaningfully perfect
    - Google Suggest Beta, Centrum, Seznam
      (http://www.google.com/webhp?complete=1&hl=en)

- Offer arranged results
  - According to synonymy class
  - Vivisimo
Discovering synonymy

• Information Sources
  – WordNet
  – monothetic clustering
  – Google API
  – NLP tools
  – Hearst Patterns
WordNet and NLP

- WordNet contains 150,000 words organized in over 115,000 synsets
- Each word has a short description *gloss*
- Key idea is to check if nouns in glosses contain the meaning (i.e. superclass) of the word.

- NLP
  - Stop words list and POS tagger
Hearst Patterns

- lexico-syntactic patterns indicating the existence of class/subclass relation in unstructured data sources

- $NP_0$ such as $NP_1, NP_2, \ldots, NP_{n-1}$ (and | or) $NP_n$
- such $NP_0$ as $NP_1, NP_2, \ldots, NP_{n-1}$ (and | or) $NP_n$
- $NP_1, NP_2, \ldots, NP_{n-1}$ (and | or) other $NP_0$
- $NP_0$ (including—especially) $NP_1, NP_2, \ldots, NP_{n-1}$ (and | or) $NP_n$
- and very common "$NP_i$ is a $NP_0$"
• get all synonyms and their glosses for given word from WordNet
• Process all glosses with NLP tools to get *candidate nouns*
• Apply “is a” pattern on each of candidate nouns and count the number of results
• Validate the results with “and other” pattern
• Compare the results and give the most probable *hypernym*
Extracting “candidate nouns”

• Considering the proper noun PLUTO
  - 3 meanings in WordNet (planet, God, cartoon)

  Glosses
  – a small planet and the farthest known planet from the sun; has the most elliptical orbit of all the planets
  – (Greek mythology) the god of the underworld in ancient mythology; brother of Zeus and husband of Persephone
  – a cartoon character created by Walt Disney
Extracting “candidate nouns”

• Considering the proper noun PLUTO
  - 3 meanings in WordNet (planet, God, cartoon)

Glosses
  - a small **planet** and the farthest known **planet** from the **sun**; has the most elliptical **orbit** of all the **planets**
  - (Greek mythology) the **god** of the **underworld** in ancient **mythology**; brother of **Zeus** and husband of **Persephone**
  - a **cartoon character** created by **Walt Disney**

  – Candidate nouns
    – planet; sun; orbit; planets;
    – Greek; god; underworld; mythology; brother; Zeus; husband; Persephone;
    – cartoon; character; Walt; Disney;
• Considering the proper noun PLUTO
  - 3 meanings in WordNet (planet, God, cartoon)

Glosses
  – a small **planet** and the farthest known **planet** from the **sun**; has the most elliptical **orbit** of all the **planets**
  – (Greek mythology) the **god** of the **underworld** in ancient **mythology**; brother of **Zeus** and husband of **Persephone**
  – a **cartoon** **character** created by **Walt Disney**

  – Candidate nouns
    – planet; sun; orbit; planets;
    – Greek; god; underworld; mythology; brother; Zeus; husband; Persephone;
    – cartoon; character; Walt; Disney;
Applying Two Patterns

• “Pluto is a planet” (1550), “Pluto is planet” (145)
• “Pluto is a sun” (2), “Pluto is sun” (0)
• “Pluto is a orbit” (0), “Pluto is orbit” (1)
• “Pluto is a planets” (0), “Pluto is planets” (0)

For each of 16 candidate nouns create similar pattern and search for it

Validation

• “Pluto and other planets” (57)
• “Pluto and other planet” (0)
• “Pluto and other suns” (0)
• “Pluto and other sun” (0)

Candidate nouns

planet; sun; orbit; planets;

• Problem with adjectives
• Plural problem
Applying Two Patterns

- "Pluto is a planet" (1550), "Pluto is planet" (145)
- "Pluto is a sun" (2), "Pluto is sun" (0)
- "Pluto is a orbit" (0), "Pluto is orbit" (1)
- "Pluto is a planets" (0), "Pluto is planets" (0)

For each of 16 candidate nouns create similar pattern and search for it

Validation

- "Pluto and other planets" (57)
- "Pluto and other planet" (0)
- "Pluto and other suns" (0)
- "Pluto and other sun" (0)

- Problem with adjectives
- Plural problem
Results

- Test set: 50 nouns from travel, space and zodiac

Table 1. Overall precision

<table>
<thead>
<tr>
<th>Total number of words in list</th>
<th>50 (100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words listed in WordNet</td>
<td>48 (96%)</td>
</tr>
<tr>
<td>Correct</td>
<td>39 (78%)</td>
</tr>
<tr>
<td>- completely correct</td>
<td>31 (62%)</td>
</tr>
<tr>
<td>- partially correct</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>Wrong</td>
<td>9 (18%)</td>
</tr>
</tbody>
</table>
Results

• Bad results

Table 2. Statistics of wrongly discovered terms

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of wrong instances</td>
<td>17</td>
</tr>
<tr>
<td>Both patterns wrong</td>
<td>7</td>
</tr>
<tr>
<td>”is a” correct, ”and other” wrong</td>
<td>4</td>
</tr>
<tr>
<td>”is a” wrong, ”and other” correct</td>
<td>6</td>
</tr>
</tbody>
</table>
Drawbacks and Future

• Drawbacks
  – No proximity search in Google
  – Slow response of Google API
  – Lot of queries to discover one meaning

• Future:
  – Add another validation step
  – Discover collocations in results
  – Discover Named Entities
Questions?

Thank you for your attention