Automatic Acquisition of Sense Examples using ExRetriever

Montse Cuadros, Jordi Atserias, Mauro Castillo, German Rigau

February 9, 2005

http://www.lsi.upc.es/~nlp/meaning

Montse Cuadros
Outline

• Automatic Acquisition of Examples for WSD
• ExRetriever: A Sense Example Retriever Tool
• The Query Language
• Examples
• WSC measure
• Experiments and Results
• Conclusions and Future Work
Outline

- **Automatic Acquisition of Examples for WSD**
- **ExRetriever: A Sense Example Retriever Tool**
- **The Query Language**
- **Examples**
- **WSC measure**
- **Experiments and Results**
- **Conclusions and Future Work**
Automatic Acquisition of Examples for WSD

- Current research on WSD uses semantically annotated corpora to train Machine Learning algorithms to WSD.

- Recent work is focusing on reducing the acquisition cost and the need for supervision in corpus-based methods for WSD.

- [Leacock et al. 98], [Mihalcea & Moldovan 99] and [Agirre & Martinez 00] automatically generate arbitrarily large corpora for unsupervised WSD training, using the knowledge contained in WordNet to formulate search engine queries over large text collections or the Web.
Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work
ExRetriever: A Sense Example Retriever Tool

- ExRetriever characterises each sense of a word as a specific query.

- Using a query construction strategy, defined *a priori*.

- Strategies can take into account the information related to words from lexical database in order to automatically generate the set of queries.

- The resulting specific queries are used to retrieve particular sense examples from a large text collection.
ExRetriever: A Sense Example Retriever Tool (2)

- Able to use different lexical databases e.g. the Multilingual Central Repository of MEANING [Atserias et al. 04]

- Different corpora (SemCor, BNC, the Web, etc.)

- ExRetriever has been powered with a declarative language to define different query construction strategies.

- Postprocess module (e.g. tagging, lemmatizing, recognizing WordNet multiwords). [Arranz et al. 04]
ExRetriever

Lexical Database

MCR
WordNet
...

(word AND
(seat OR person
OR support OR
back))...

a seat for one person,
with a support for the back

It contained a desk,
files, a typewriter on a
stand, and two big
leather armchair

.word sense 1
word sense 2
....
word sense N

query

Retriever
Process

Sense
Examples

ExRetriever Tool

corpus

BNC SemCor
www
...

Query
Strategies

Evaluation

ESTADISTICS

Montse Cuadros
Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work
The Query Language

- **Operators** "and", "or" and "not".

- **Constants:**
  - `noempty` a parameter of `Glos` function to remove empty words.
  - `senses` lemma#POS#sense number (e.g. church#n#2)
  - `relations` names of the lexical relationships used as parameters to "rel" and "nrel" (e.g. hyponym).

- **Functions** Currently,
  - `Glos` to build expressions from the words in the gloss.
  - `rel` to look up the different relations in the Lexical database
  - `nrel` similar to `rel`, establishes the maximum polysemy of the returned senses.
Outline

• Automatic Acquisition of Examples for WSD

• ExRetriever: A Sense Example Retriever Tool

• The Query Language

• Examples

• WSC measure

• Experiments and Results

• Conclusions and Future Work
Example

**Meaning1Semcor:** \( \text{Glos(or, and, noempty)} \) OR \( \text{or(nrel(1,syns))} \) OR \( \text{or(nrel(1,hypo))} \)

The first function \( \text{Glos(or, and, noempty)} \) returns a logical formula which is the target word (e.g. *chair*) and the union set with *or* of the *noempty* words of the *gloss* of the sense considered (e.g. char#n#3).

The second function, \( \text{or(nrel(1,syns))} \) returns the union set with *or* of the monosemous synonyms.

Finally, \( \text{or(nrel(1,hypo))} \) returns the union set of the monosemous hyponyms.
Example: the noun *Chair*

<table>
<thead>
<tr>
<th>sense</th>
<th>gloss</th>
<th>hypo</th>
<th>syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>n#1</td>
<td><em>a seat for one person</em>, <em>with a support for the back</em></td>
<td>armchair (2)</td>
<td>barber_chair ...</td>
</tr>
</tbody>
</table>

Table 1: Queries for *chair* noun using **Meaning1SemCor**

Table 2: Sense of *chair* noun in wordNet 1.6
Sense II

<table>
<thead>
<tr>
<th>char#n#2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(chair AND (position OR professor)) OR (professorship)</td>
</tr>
</tbody>
</table>

Table 3: Queries for chair noun using Meaning1SemCor

<table>
<thead>
<tr>
<th>sense</th>
<th>gloss</th>
<th>hypo</th>
<th>syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>n#2</td>
<td>the position of professor</td>
<td></td>
<td>professorship</td>
</tr>
</tbody>
</table>

Table 4: Sense of chair noun in wordNet 1.6
Sense III

<table>
<thead>
<tr>
<th>sense</th>
<th>gloss</th>
<th>hypo</th>
<th>syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>n#3</td>
<td>the officer who presides at the meetings of an organization</td>
<td>vice_chairman</td>
<td>president (6)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chairman</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chairwoman</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chairperson</td>
</tr>
</tbody>
</table>

Table 5: Queries for *chair* noun using **Meaning1SemCor**

Table 6: Sense of *chair* noun in wordNet 1.6
Sense IV

<table>
<thead>
<tr>
<th>sense</th>
<th>gloss</th>
<th>hypo</th>
<th>syn</th>
</tr>
</thead>
</table>
| n#4  | *an instrument of death by electrocution that resembles a chair* | | electric_chair  
 | | | death_chair  
 | | | hot_seat |

Table 7: Queries for *chair* noun using **Meaning1SemCor**

Table 8: Sense of *chair* noun in wordNet 1.6
Example of the Post Processing

<Example Sentences="1" src="brownv/tagfiles/br-a10#43160" > Seeking this two-year <MEANING synsetPOS="v" baseSense="1" baseLema="call" origPOS="n" rel="hypo" synsetSense="1" synsetLema="term" basePOS="v" > term </MEANING> are James_Culbertson, Dwight_M._Steeves, James_C._Piersee, W._M._Sexton and Theodore_W._Heitschmidt. </Example>

lemaTAG: term  lemaORIG: term
posTAG: n  posORIG: v
Example of the Post Processing

Since the sides are also covered up to the spray, they are also rough sanded in that area.

lemaTAG: rail  lemaORIG: rails
posTAG: n     posORIG: n
Example of an extracted sentence

<Example Sentences="1" src="brown2/tagfiles/br-l15#104577"> It contained a desk, files, a typewriter on a stand, and two big leather armchairs. </Example>

Results for chair against SemCor

<table>
<thead>
<tr>
<th>Sense</th>
<th>Ok</th>
<th>Ko</th>
<th>NoTag</th>
<th>#Sense</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>n#1</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>34</td>
<td>89</td>
<td>44</td>
<td>59</td>
</tr>
<tr>
<td>n#2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>50</td>
<td>25</td>
<td>33</td>
</tr>
<tr>
<td>n#3</td>
<td>7</td>
<td>0</td>
<td>32</td>
<td>11</td>
<td>100</td>
<td>64</td>
<td>78</td>
</tr>
<tr>
<td>n#4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Totals</td>
<td>24</td>
<td>4</td>
<td>33</td>
<td>48</td>
<td>86</td>
<td>24</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 9: Results of chair#n applying Meaning1SemCor
Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work
WSC measure

- Precision, Recall and F1 don’t show if the examples retrieved cover all the sense of a word.

- This is a crucial issue if we want to use the acquired examples to train supervised WSD systems.

- We have defined a new measure, WSC (word sense coverage).

WSC measure:

\[
WSC = 100 \sum_{w=1}^{n} \times \frac{SensesWithinRetrievedExamples(w)}{SensesWithinCorpus(w)}
\]
Outline

• Automatic Acquisition of Examples for WSD
• ExRetriever: A Sense Example Retriever Tool
• The Query Language
• Examples
• WSC measure
• Experiments and Results
• Conclusions and Future Work
Experiments and Results

- MCR as lexical Database.
- Semcor as corpus.
- 6 different query construction strategies.
- Precision, Recall, F1 and WSC measure.

<table>
<thead>
<tr>
<th>Q</th>
<th>Ok</th>
<th>Ko</th>
<th>NoTag</th>
<th>#Sense</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>WSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lea1</td>
<td>851</td>
<td>10</td>
<td>371</td>
<td>23254</td>
<td>98,84</td>
<td>3,66</td>
<td>7,06</td>
<td>23</td>
</tr>
<tr>
<td>Mol1</td>
<td>153</td>
<td>1</td>
<td>83</td>
<td>3241</td>
<td>99,35</td>
<td>4,72</td>
<td>9,01</td>
<td>10</td>
</tr>
<tr>
<td>Mol3</td>
<td>1987</td>
<td>22474</td>
<td>1303</td>
<td>7611</td>
<td>8,12</td>
<td>26,11</td>
<td>12,39</td>
<td>47</td>
</tr>
<tr>
<td>Mea1</td>
<td>2314</td>
<td>22617</td>
<td>1415</td>
<td>9490</td>
<td>9,28</td>
<td>24,38</td>
<td>13,44</td>
<td>54</td>
</tr>
<tr>
<td>Mea2</td>
<td>4513</td>
<td>37688</td>
<td>2986</td>
<td>17171</td>
<td>10,69</td>
<td>26,28</td>
<td>15,20</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 10: Overall figures
Experiments and Results (2)

• **Moldo1** and **Lea1** strategies obtain the best precision (around 99%), but poor coverage and WSC.

• **Meaning1**, **Meaning2**, **Moldo3** methods obtain much better recall (about 25% vs 5%) and WSC but less precision.

• **Meaning2**, the best WSC obtaining examples for 58% of the senses.

• **Moldo2** strategy do not provide results in SemCor, as it looks for the complete synset gloss.
Outline

- Automatic Acquisition of Examples for WSD
- ExRetriever: A Sense Example Retriever Tool
- The Query Language
- Examples
- WSC measure
- Experiments and Results
- Conclusions and Future Work
Conclusions

- ExRetriever, a query-based system to extract sense examples from corpus has been described.

- Some preliminar experiments have been presented. They have been used to evaluate the performance of different types of query construction strategies.

- Using ExRetriever, new strategies can be easily defined, executed and evaluated.
Future Work

• Experiment other strategies. (e.g. performing full parsing on the glosses could help discarding irrelevant words from glosses).

• Using the knowledge already contained into the MCR (e.g., selectional preferences, domain information, etc.) to better model sense words as queries.

• Use alternative schemata for building queries, such as the incremental process performed by [Leacock et al. 98].

• Follow [Widdows 03]. It seems that most of the errors produced because of the substitution of the target word for their relatives can be avoided.
Future Work (2)

- Use other sense tagged corpora for direct comparisons of ExRetriever (e.g. DSO).

- Perform indirect evaluations using supervised WSD systems on the acquired sense examples.

- Once acquired a sense tagged corpus using ExRetriever, we will use several Machine Learning algorithms to perform several cross-comparisons with respect to other sense tagged resources (SemCor, DSO and those resources provided by Senseval).
Available here:

You can download it here:

http://www.lsi.upc.es/~nlp/meaning/downloads.html
Thanks for your attention

http://www.lsi.upc.es/~nlp/meaning

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANING IST-2001-34460).
References


ExRetriever


Montse Cuadros