Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation

Eneko Agirre and Aitor Soroa
<a.soroa@ehu.es>

University of the Basque Country

LREC, Marrakesh 2008
Introduction

- WSD: assign a sense to a word in a particular context
- Supervised WSD performs best
  - but needs large amounts of hand-tagged data
- Knowledge-based WSD
  - Exploit information present on a LKB
  - No further corpus evidence
Knowledge-based WSD

- **Traditional approach:**
  - Assign a sense to an ambiguous word by comparing each of its senses with those of the surrounding context
  - Some semantic similarity metric used for calculating the relatedness among senses
  - Due to combinatorial explosion, words are disambiguated individually

- **Graph based methods**
  - Graph-based techniques to exploit the structural properties of the graph underlying the LKB
  - Find globally optimal solutions given the relations between entities
  - Disambiguate large portions of text in one go
Main goal of the work

- Novel graph-based method for performing unsupervised WSD
- The method is independent of underlying LKB
  - Applied to Multilingual Central Repository (MCR)
- Evaluate separate and combined performance of several relation types of the MCR
Outline

1. Introduction
2. A graph algorithm for knowledge-based WSD
3. Multilingual Central Repository
4. Experiments
5. Conclusions
A graph algorithm for knowledge-based WSD

Represent the LKB as a graph
- Nodes are the concepts ($v_i$)
- Edges are relations among concepts ($e_{ij}$)

Given an input context
- $W_i$, $i = 1 \ldots m$: content words (nouns, verbs, adjectives and adverbs)
- $\text{Synsets}_i = \{v_{i1}, \ldots, v_{in}\}$: synsets associated to word $i$

Two steps for WSD
1. Extract a representative subgraph: disambiguation subgraph
2. Find the “best” synsets of the subgraph
Extracting the disambiguation subgraph

Subgraph extraction:
- For each word $W_i$, $i = 1 \ldots m$
- For each synset $v_{i1} \ldots v_{in}$ input word $W_i$
  - Find the shortest paths from $v_{ij}$ to synsets of rest of words (BFS search)
  - Create subgraph by joining all minimum distance paths

The vertices and relations of the subgraph are particularly relevant for a given input context.
Identifying the best synsets: PageRank

- Google’s PageRank (Brin and Page, 1998): model a random walk on the graph
  - A walker takes random steps
  - Converges to a stationary distribution of probabilities
- \( G = (V, E) \) a graph
  - \( \text{In}(V_i) \) = nodes pointing to \( V_i \)
  - \( d_j = \text{degree of node } v_j \)

\[
\text{PR}(V_i) = (1 - \alpha) + \alpha \sum_{j \in \text{In}(V_i)} \frac{1}{d_j} \text{PR}(V_j)
\]

Usually \( \alpha = 0.85 \). Models random jumps.
Identifying the best synsets: PageRank

- PageRank ranks vertices according to their structural importance on the graph
- Apply PageRank over disambiguation subgraph
- Select the synsets with maximum rank for each input word
  - In case of ties, select all synsets with same rank
Multilingual Central Repository (MCR)

- Knowledge base built within the MEANING project
  - Multilingual interface for integrating and distributing all the knowledge acquired in the project
- Current version: 1,500,000 relations
  - Most of them automatic
- MCR integrates
  - ILI based on WN1.6
  - EWN Base Concepts
  - MultiWordNet Domains (MWND)
  - Local WordNets connected to the ILI
    - English WN1.5, 1.6, 1.7, 1.7.1
    - Basque, Catalan, Italian and Spanish WordNets
  - Semantic preferences
    - Acquired automatically from Semcor and BNC
  - eXtended WordNet
  - Instances, including named entities
In this work, we have used:
- WN1.6: English WordNet 1.6 synsets and relations
- WN2.0: English WordNet 2.0 relations (mapped to WN1.6 synsets)
- XNET: eXtended WordNet (gold, silver and normal)
- sPref: Selectional preferences
- sCooc: Coocurrence
- WN1.7: English WordNet 1.7 synsets and relations

sPref and sCooc extracted from Semcor
- system benefits from supervised information when using these
We have tried different set of relations

<table>
<thead>
<tr>
<th>Name</th>
<th>Relations</th>
<th>#synsets</th>
<th>#relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>M16</td>
<td>WN1.6, REL2.0, XNET, sPref, sCooc</td>
<td>99,634</td>
<td>1,651,445</td>
</tr>
<tr>
<td>M16_wout_sPref</td>
<td>WN1.6, REL2.0, XNET, sCooc</td>
<td>99,634</td>
<td>1,519,833</td>
</tr>
<tr>
<td>M16_wout_sCooc</td>
<td>WN1.6, REL2.0, XNET, sPref</td>
<td>99,632</td>
<td>798,453</td>
</tr>
<tr>
<td>M16_wout_Xnet</td>
<td>WN1.6, REL2.0, sPref, sCooc</td>
<td>99,238</td>
<td>1,169,300</td>
</tr>
<tr>
<td>M16_wout_Semcor</td>
<td>WN1.6, REL2.0, XNET</td>
<td>99,632</td>
<td>637,290</td>
</tr>
<tr>
<td>M17</td>
<td>WN1.7, XNET</td>
<td>109,359</td>
<td>620,396</td>
</tr>
<tr>
<td>M16_wout_WXnet</td>
<td>sPref, sCooc</td>
<td>27,336</td>
<td>1,024,698</td>
</tr>
</tbody>
</table>

Two main groups

- **M16**: Based on WordNet 1.6
- **M17**: Based on WordNet 1.7
Outline

1. Introduction
2. A graph algorithm for knowledge-based WSD
3. Multilingual Central Repository
4. Experiments
5. Conclusions
Experiment setting

- Applied to Senseval 3 All Words dataset
  - Based on WordNet 1.7
- Contexts of at least 20 words
  - Adding sentences immediately before and after
Experiment results

<table>
<thead>
<tr>
<th>Relations</th>
<th>All</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj.</th>
<th>Adv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M16</td>
<td>57.30</td>
<td>62.30</td>
<td>49.00</td>
<td><strong>62.40</strong></td>
<td>92.90</td>
</tr>
<tr>
<td>M16_wout_sPref</td>
<td>57.90</td>
<td><strong>63.10</strong></td>
<td><strong>49.80</strong></td>
<td>61.80</td>
<td>92.90</td>
</tr>
<tr>
<td>M16_wout_sCooc</td>
<td>53.00</td>
<td>58.10</td>
<td>44.20</td>
<td>58.30</td>
<td>92.90</td>
</tr>
<tr>
<td>M16_wout_Xnet</td>
<td>57.60</td>
<td><strong>63.10</strong></td>
<td>49.60</td>
<td>61.00</td>
<td>92.90</td>
</tr>
<tr>
<td>M16_wout_WXnet</td>
<td>55.30</td>
<td>58.70</td>
<td>48.70</td>
<td>60.80</td>
<td><strong>85.70</strong></td>
</tr>
<tr>
<td>Unsupervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M16_wout_semcor</td>
<td>53.70</td>
<td>59.50</td>
<td>45.00</td>
<td>57.80</td>
<td>92.90</td>
</tr>
<tr>
<td>M17</td>
<td><strong>56.20</strong></td>
<td><strong>61.60</strong></td>
<td><strong>47.30</strong></td>
<td><strong>61.80</strong></td>
<td>92.90</td>
</tr>
</tbody>
</table>

- Supervised relations achieve best overall results
  - Specially sCooc, not so with sPref
  - Using only supervised also yields good results

- Unsupervised results: M17 performs best
  - probably due to mapping noise
## Comparison to related work

<table>
<thead>
<tr>
<th>System</th>
<th>All</th>
<th>Noun</th>
<th>Verb</th>
<th>Adj.</th>
<th>Adv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mih05</td>
<td>52.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sin07</td>
<td>52.4</td>
<td>60.45</td>
<td>40.57</td>
<td>54.14</td>
<td>100</td>
</tr>
<tr>
<td>Nav07</td>
<td>-</td>
<td>61.9</td>
<td>36.1</td>
<td>62.8</td>
<td>-</td>
</tr>
<tr>
<td>M17</td>
<td>56.20</td>
<td>61.60</td>
<td>47.30</td>
<td>61.80</td>
<td>92.90</td>
</tr>
<tr>
<td>MFS</td>
<td>60.9-62.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GAMBL</td>
<td>65.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- **Mih05, Sin07**: create a complete weighted graph with synsets of the words in the input context. Weights calculated with similarity measures. Apply PageRank for disambiguating.
- **Nav07**: create subgraph of LKB using DFS search. LKB: Manually enriched WordNet.
Outline

1. Introduction
2. A graph algorithm for knowledge-based WSD
3. Multilingual Central Repository
4. Experiments
5. Conclusions
Conclusions

- Graph-based method for performing knowledge-based WSD
- Exploits the structural properties of the graph underlying the chosen knowledge base
- The method is not tied to any particular knowledge base
- Evaluation performed on Senseval-3 All Words
- Evaluation of separate and combined performance of each type of relation in the MCR
  - Validate the contents of the MCR and their potential for WSD
- MCR valuable for performing WSD
  - Relations coming from hand-tagged corpora are the most valuable
- Version of WordNet is highly relevant
- Our graph-based WSD system is competitive with the current state-of-the-art
  - Yields best results that can be obtained using publicly available data