DutchSemCor: Targeting the ideal sense-tagged corpus

Piek Vossen¹, Attila Görög¹, Rubén Izquierdo², Antal van den Bosch²

1 VU University Amsterdam. Amsterdam. The Netherlands
2 Tilburg University. Tilburg. The Netherlands

Build a sense-tagged corpus for Dutch

- Represent all senses of words
- Represent the variety of contexts
- Provide information on the sense distribution
- 3000 most frequent Dutch words (also most polysemous)
- 100 examples per sense
- 3-4 avg. senses per word --> 1 million tokens

Sense tagged corpus

<table>
<thead>
<tr>
<th>Annotation method</th>
<th>Textual coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential tagging</td>
<td>All - words</td>
</tr>
<tr>
<td>Targeted tagging</td>
<td>Lexical sample</td>
</tr>
<tr>
<td></td>
<td>Balanced-sense</td>
</tr>
<tr>
<td></td>
<td>Balanced-context</td>
</tr>
<tr>
<td></td>
<td>Sense-Probability</td>
</tr>
</tbody>
</table>

Project Methodology. 3 Phases

1) Human manual annotation

Goal: Balanced-sense corpus

- Double annotation for each token
- 282,503 tokens
- Very clear and good examples selected

2) Active Learning

Goal: Balanced-context corpus

- Only lemmas performing < 80 in accuracy are processed
- 50 examples per sense according to:
  - TIMBL confidence
  - Distance to the nearest neighbor
  - Low Distance (LD): similar examples
  - High Distance (HD): different examples

3) Clustering

Goal: sense-probability corpus

- Similar to Word Sense Induction
- Clustering techniques different to WSD
- Cluster remain not tagged SONAR to:
  - Discover new senses
  - Use annotated instances to discover clusters with a predominant word sense and automatically tag the cluster

A 500-million-token corpus is not big enough to create a balanced-sense corpus!!