Automatic word-sense disambiguation for Dutch using dependency knowledge

› Hessel Haagsma

› Supervision: Gertjan van Noord
Task

“Foolen kondigt de geboorte aan van drie nieuwe tijdschriften.”
“Foolen announced the birth of three new journals.”

What is the meaning of 'geboorte'?
1. 'het baren van een kind' – childbirth
2. 'afkomst, oorsprong' - origin, descent, ancestry
3. 'ontstaan, oprichting' - foundation, creation
Intuition

(1) “Two occurrences of the same word have identical meanings if they have similar local contexts.”

(2) “Two different words are likely to have similar meanings if they occur in identical local contexts.”

Implementation for Dutch

Which resources are needed to implement this for Dutch?

› Local context database, i.e. a large, dependency-parsed corpus
› Semantic database and ontology, for sense selection and similarity measures
› Sense-annotated corpus, for testing and frequency estimation
Local context database

› Dependency-parsed corpus: LassyLarge
› Local context: *van-mod-geboorte-head*

› Version 1: use DFS-script for all triples containing only nouns, adjectives, adverbs, verbs
› Version 2: use Alpino-generated triples: adds prepositions, indirect relations
Semantic database

- Semantic database: Cornetto
- Retrieve possible senses
- Semantic similarity measure:

\[
sim(hill, coast) = \frac{2 \times \log(GeoForm)}{\log(hill) + \log(coast)}
\]
Sense-annotated corpus

- Sense-annotated corpus: DutchSemCor
- 1M annotations, 250K manually done
- Sense inventory maps to Cornetto
- Minimum number of instances of each sense
- 942 word sample as test set: 564 nouns, 327 verbs, 51 adjectives
**Results (1)**

<table>
<thead>
<tr>
<th>Local context database</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-sense baseline</td>
<td>28.24</td>
</tr>
<tr>
<td>Version 1 (script)</td>
<td>31.32</td>
</tr>
<tr>
<td>Version 2 (Alpino)</td>
<td>33.76</td>
</tr>
<tr>
<td>With manual part of DSC</td>
<td>36.41</td>
</tr>
<tr>
<td>40 similar words</td>
<td>37.15</td>
</tr>
</tbody>
</table>
Results (2)

- Nouns: 40.58%, verbs: 36.42%, adjectives: 23.40%
- Adjectives do not fit in hierarchy
- Lin (1997): 56.1%, baseline 58.9%. Nouns only
- Corpus-dependent baseline
Problems

› Similar words not sufficient for differentiation
› Frequency-balanced corpus distorts semantic similarity
› No context: 'Zalig!'
› Uninformative contexts: van - mod, ben - su
› Obscure context: X - internetpretpost - obj1
Improvement

› Implement named-entity placeholders
› Use sense-annotated corpus with natural sense frequencies
› Exploit other relations in semantic database: (near-)synonym, (near-)antonym
› Make local contexts larger
› Explore alternative similarity measures