A Coreference Corpus and Resolution System for Dutch

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Véronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, Jean-Luc Verschelde, Frederik Coppens
COREA project: Coreference Resolution for Extracting Answers

URL: http://www.cnts.ua.ac.be/~iris/corea.html

Team:

- **University of Antwerp**: Walter Daelemans, Iris Hendrickx, Véronique Hoste
- **University Groningen**: Gosse Bouma, Anne-Marie Mineur, Geert Kloosterman
- **Language & Computing N.V.**: Jean-Luc Verschelde, Frederik Coppens, Joeri Van Der Vloet
Overview of the talk

- Corea project
- Corpus and annotation
- Coreference resolution module
- Evaluation
  - Effect on Question Answering
  - Effect on Information Extraction
Application-oriented approach

Many Natural Language Processing applications such as Information Extraction and Automatic Summarization require accurate identification of coreference relations between noun phrases.

Gas station collapses

Gas station Hoezaar next to highway A58 has collapsed monday afternoon. The building came down after being hit by a truck with a flat tyre.
COREA Goals

- Annotation guideline manual for Dutch
- Annotated evaluation corpus of 100k words
- Coreference resolution tool
- Integration and evaluation of tool in NLP application, Information Extraction and Question answering
Annotation

- Coreference is restricted to names, pronouns, noun phrases (NP).
- 200K words
- Different text genres: newspaper, spoken language, medical domain, Dutch and Flemish
- Different types of coreference relations
Types of Coreference

• Identity (IDENT)

* Xavier Malisse qualified for the semi finals in Wimbledon. The Flemish tennis player will play against an unknown opponent.

• Quantification (BOUND)

* Everybody did what they could.

• Superset – Subset (BRIDGE)

* 200 people died in that plain crash. Forty-six are buried here on this cemetery.

• Predicative relations (PRED)

* Michel Beuter is a writer.

• Special cases: negation, modality, time dependency
## Corpus statistics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>DCOI</th>
<th>CGN</th>
<th>MedEnc</th>
<th>Knack</th>
</tr>
</thead>
<tbody>
<tr>
<td>#docs</td>
<td>105</td>
<td>264</td>
<td>497</td>
<td>267</td>
</tr>
<tr>
<td>#tokens</td>
<td>35,166</td>
<td>33,048</td>
<td>135,828</td>
<td>122,960</td>
</tr>
<tr>
<td>#ident</td>
<td>2,888</td>
<td>3,334</td>
<td>4,910</td>
<td>9,179</td>
</tr>
<tr>
<td>#bridge</td>
<td>310</td>
<td>649</td>
<td>1,772</td>
<td>na</td>
</tr>
<tr>
<td>#pred</td>
<td>180</td>
<td>199</td>
<td>289</td>
<td>na</td>
</tr>
<tr>
<td>#bound</td>
<td>34</td>
<td>15</td>
<td>19</td>
<td>43</td>
</tr>
</tbody>
</table>
Inter-annotator Agreement

Experiment:
2 annotators, 29 documents, +- 500 relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ident</td>
<td>76%</td>
</tr>
<tr>
<td>Bridge</td>
<td>33%</td>
</tr>
<tr>
<td>Pred</td>
<td>56%</td>
</tr>
<tr>
<td>Bound</td>
<td>0 %</td>
</tr>
</tbody>
</table>
Visualization

[Image: Visualization page from Marrakech, LREC 2008]
Coreference resolution as classification task

Supervised Machine Learning approach

- Identify the NPs in the text,
- Link every NP to the previous NPs,
- Step one: classify each pair as coreferential or not
- Step two: make coreference chain of positive pairs
Effect on Question Answering

Evaluation Dutch QA system Joost:

**The Fact Extractor**: extracts answers to frequent questions off-line, based on manually developed patterns

*Who was born when?*

*Which city is the capital of which country?*

Example

Fact type: *What number_of_inhabitants for Location?*

Sentence: *The village has 10,000 inhabitants*

→ resolve antecedent of the village to extract the fact

Marrakech, LREC 2008
Effect on Question Answering

Coreference information (rules-based) in Fact Extractor
More facts are extracted: from 93K to 145K

How many questions are answered correctly?

<table>
<thead>
<tr>
<th>variant</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>without</td>
<td>65.0%</td>
</tr>
<tr>
<td>with</td>
<td>70.0%</td>
</tr>
</tbody>
</table>

Table 1: Number of correctly answered questions in QA@CLEF 2005 test set.
Effect on Information Extraction

Relation Finder predicting medical semantic relations. Based on Spectrum Medical Encyclopedia annotated with medical concepts and relations between them\(^a\)

Medical concepts: con\_disease, con\_person, con\_treatment

Relations: rel\_is\_symptom\_of, is\_cause\_of, rel\_treats

\(^a\)Corpus developed in IMIX Rolaquad project
Relation Finder

- Core: Maximum Entropy Modeling algorithm
- Trained on 2000 encyclopedia entries
- Tested on two test sets of 50 and 500 different entries
- Evaluated with and without coreference information as predicted by our module
Effect on Information Extraction

Results with and without coreference information:

<table>
<thead>
<tr>
<th>test set</th>
<th>without</th>
<th>with</th>
</tr>
</thead>
<tbody>
<tr>
<td>small(50)</td>
<td>53.03</td>
<td>53.51</td>
</tr>
<tr>
<td>Big(500)</td>
<td>59.15</td>
<td>59.60</td>
</tr>
</tbody>
</table>

Table 2: F-Scores of Relation Finder.
Conclusions

• Current results show a marginal but positive effect

• More work is needed to refine our approach
Future Plans

• Groningen: Improving the coreference resolution module in QA system JOOST

• Antwerpen: DEASO project: multi-document summarization
Thanks for your attention.