# A Coreference Corpus and Resolution System for Dutch
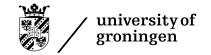
Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans,Véronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, Jean-Luc Verschelde, Frederik Coppens

Universiteit
Antwerpen

L&C
Language & Computing

university of
groningen

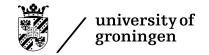# COREA project: Coreference Resolution for Extracting Answers

*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
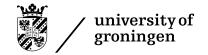
# Overview of the talk

- Corea project

- Corpus and annotation

- Coreference resolution module

- Evaluation

    – Effect on Question Answering

    – Effect on Information Extraction
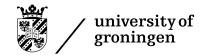
# Application-oriented approach

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.
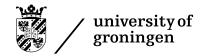
# COREA Goals

- 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

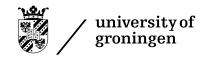- Different types of coreference relations

# 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

| Corpus | DCOI | CGN | MedEnc | Knack |
|---|---|---|---|---|
| #docs | 105 | 264 | 497 | 267 |
| #tokens | 35,166 | 33,048 | 135,828 | 122,960 |
| #ident | 2,888 | 3,334 | 4,910 | 9,179 |
| #bridge | 310 | 649 | 1,772 | na |
| #pred | 180 | 199 | 289 | na |
| #bound | 34 | 15 | 19 | 43 |

# Inter-annotator Agreement

Experiment:

2 annotators, 29 documents, +- 500 relations

| Relation | F-score |
|----------|---------|
| Ident | 76% |
| Bridge | 33% |
| Pred | 56% |
| Bound | 0 % |

# 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 <span style="color:red">the village</span> to extract the fact

# 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?

| variant | accuracy |
|---------|----------|
| without | 65.0% |
| with | 70.0% |

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
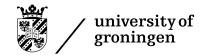medical concepts and relations between them[a]

Medical concepts: *con_disease, con_person, con_treatment*
Relations: *rel_is_symptom_of, is_cause_of, rel_treats*

---

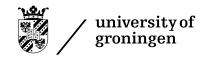[a]Corpus developed in IMIX Rolaquad project

# 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:

| test set | without | with |
|---|---|---|
| small(50) | 53.03 | 53.51 |
| Big(500) | 59.15 | 59.60 |

Table 2: F-Scores of Relation Finder.

# Conclusions

- Current results show a marginal but positive effect

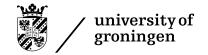- More work is needed to refine our approach

# Future Plans

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

- Antwerpen: DEASO project: multi-document summarization

# Thanks for your attention.