A Coreference Corpus and Resolution System for Dutch

Iris Hendrickx*, Gosse Bouma‡, Frederik Coppens◇, Walter Daelemans*, Veronique Hoste*
Geert Kloosterman‡, Anne-Marie Mineur‡, Joeri Van Der Vloet◇, Jean-Luc Verschelde◇

*CNTS, University of Antwerp,
Prinsstraat 13, 2000 Antwerpen, Belgium
{iris.hendrickx, walter.daelemans, veronique.hoste} @ua.ac.be
‡Information Science, University of Groningen,
Groningen, The Netherlands
{g.bouma, g.j.kloosterman, a.m.c.mineur}@rug.nl
◇Language and Computing NV,
Kortrijksesteenweg 1038, B-9051 Sint-Denijs-Westrem, Belgium
info@landcglobal.com

Abstract

1. Introduction

Coreference resolution is a key ingredient for the automatic interpretation of text. The extensive linguistic literature on this subject has restricted itself mainly to establishing potential antecedents for pronouns. Practical applications, such as Information Extraction, summarization and Question Answering, require accurate identification of coreference relations between noun phrases in general. Currently available computational systems for assigning such relations automatically have been developed mainly for English (e.g. (Soon et al., 2001), (Harabagiu et al., 2001), (Ng and Cardie, 2002a)). A large part of these approaches are corpus-based and require the availability of a sufficient amount of annotated data. For Dutch, annotated data is scarce and coreference resolution systems are scarce (Hoste, 2005). In the COREA project we tackled these problems. We developed guidelines for the manual annotation of coreference resolution for Dutch and created a corpus annotated with coreferential relations of over 200k words.

We also developed a coreference resolution module for Dutch which we evaluated in two ways. The standard approach to evaluate a coreference resolution system is to compare the predictions of the system to a hand-annotated gold standard test set (cross-validation). A more practical oriented evaluation is to test the usefulness of coreference relation information in an NLP application. We present the results of both this application-oriented evaluation of our system and of a standard cross-validation evaluation. We ran experiments with an Information Extraction module for the medical domain, and measure the performance of this module with and without the coreference relation information. In a separate experiment we also evaluated the effect of coreference information produced by a simple rule-based coreference module in a Question Answering application.

We discuss the corpus creation process in section 2. In section 3. and **??** we present our coreference resolution application and the results of cross validation experiments. In section 4. we present an extrinsic evaluation of our coreference resolution module on the Information Extraction and our experiments for the Question Answering application.

2. Corpus annotation

Guidelines and corpus selection

For the annotation of coreference relations we developed a set of annotation guidelines largely based on the MUC-6 (Fisher et al., 1995) and MUC-7 (MUC-7, 1998) annotation scheme for English. Coreference relations are annotated using SGML tagging within the text stream. The details of our annotation scheme can be found in the COREA annotation guidelines (Bouma et al., 2007). Here we give a broad overview of the type of coreference relations annotated in our corpus.

Annotation focuses primarily on coreference or IDENTITY relations between noun phrases, where both noun phrases refer to the same extra-linguistic entity. Example 1 presents an identity relation between *Xavier Malisse* and *De Vlaamse tennisser*.

(1) [Xavier Malisse]$_1$ heeft zich geplaatst voor de halve finale in Wimbledon. [De Vlaamse tennisser]$_1$ zal dan tennissen tegen een onbekende tegenstander.
(*English: Xavier Malisse has qualified for the semi-finals at Wimbledon. The Flemish tennis player will play against an unknown opponent at that occasion.*)

We annotate several other coreference relations and flag certain special cases. We annotate BOUND relations where an anaphor refers to a quantified antecedent. An example is shown in 2.

(2) [iedereen]$_1$ heeft [zijn]$_1$ best gedaan.
*English: Everybody$_1$ did what they$_1$ could.?**

Another type of relations are superset–subset or group–member relations, which we denote with the term BRIDGE. Example 3 presents such a bridge relation in which the anaphor is a subset of the antecedent.

(3) In de Raadsvergadering is het vertrouwen opgezegd in [het college]$_1$. In een motie is gevraagd aan [alle wethouders]$_2$ hun ontslag in te dienen.

English: In the council meeting the confidence in [mayor-and-aldermen]$_1$ has been withdrawn. A motion requests that [all aldermen]$_2$ resign.

We also marked predicative relations (PRED). These are not strictly speaking coreference relations, but we annotated them for a practical reason. Such relations express extra information about the referent that can be useful for example for a question-answering application. Example 4 shows such PRED relation.

(4) [Michiel Beute]$_1$ is [schrijver]$_1$.
English: [Michiel Beute]$_1$ is [writer]$_1$.

In cases where a coreference relation is negated, modified or time dependent, the relation is annotated with a warning flag. We also mark cases in which two noun phrases point to the same referent but have a difference in their meaning. Example 5 shows such special case. The anaphor *woord (English: name)* does not refer to the same object in the real world as the antecedent, but refers to its lexical representation.

(5) [een doorstroomstrook] langs de A4 ja zoals ze 't noemen van Amsterdam naar de Belgische grens ... ook [een mooi woord] .
English: [a rush hour lane] next to the A4 as they call it from Amsterdam to the Belgian border ... also [a pretty name].

To create a annotated corpus for Dutch, we annotated texts from different sources:

- Dutch news paper articles gathered in the DCOI project [1]
- transcribed spoken language from the Corpus of Spoken Dutch (CGN)[2]
- entries from the Spectrum (Winkler Prins) medical encyclopedia as gathered in the IMIX ROLAQUAD project[3] (MedEnc)
- Already available was the KNACK-2002 corpus based on KNACK, a Flemish weekly news magazine.

For training and evaluation, we also used annotated material from the KNACK-2002 corpus (a Flemish weekly news magazine). The annotation of this corpus is described in (Hoste, 2005), and is compatible with the annotation in COREA. Note that the corpus covers a number of different genres (speech transcripts, news, medical text) and contains both Dutch and Flemish sources. The latter is particularly relevant as the use of pronouns is different in Dutch and Flemish. Table 1 presents the number of annotated identity, bridging, predicative and bound relations in the different text sources.

[1]DCOI:http://lands.let.ru.nl/projects/d-coi/

[2]CGN: http://lands.let.ru.nl/cgn/

[3]IMIX:http://ilk.uvt.nl/rolaquad/

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

Table 1: Corpus statistics for the coreference corpora used in the Corea project.

As annotation environment we used the MMAX2 annotation software.[4] For the CGN and DCOI material, manually corrected syntactic dependency structures were available. Following the approach of Hinrichs et al. (2005), we used these to create an initial set of markables and to simplify the annotation task. The labeling was done by several annotators who had a linguistic background. Due to time restrictions each document was only annotated once.

Inter-annotator agreement

To estimate the inter-annotator agreement for this task, 29 documents from CGN and DCOI were annotated independently by two annotators. These annotation statistics are given in table 2 .

Annotator	1	2
IDENT	460	397
BRIDGE	45	43
PRED	11	31
BOUND	3	3
Total	517	470

Table 2: Annotation Statistics for Annotator 1 and 2

For the IDENT relation, we compute inter-annotator agreement as the F-measure of the MUC-scores (Vilain et al., 1995) obtained by taking one annotation as 'gold standard' and the other as 'system output'. For the other relations, we compute inter-annotator agreement as the average of the percentage of *anaphor-antecedent* relations in the gold standard for which an *anaphor-antecedent'* pair exists in the system output, and where *antecedent* and *antecedent'* belong to the same cluster (w.r.t. the IDENT relation) in the gold standard. Inter-annotator agreement for IDENT is 0.76 (f-score), for bridging is 33% and for PRED is 56%. There was no agreement on the (small number of) BOUND relations. The agreement score for IDENT is comparable, though slightly lower, than those reported for comparable tasks for English and German (Hirschman et al., 1997; Versley, 2006). Poesio and Vieira (1998) report 59% agreement on annotating 'associative coreferent' definite NPs, a relation comparable to our BRIDGE relation.

[4]MMAX2 is available at: http://www.eml-research.de

The main sources of disagreement were

1. Cases where an annotator fails to annotate a coreference relation.
2. Cases where a BRIDGE or PRED relation is annotated as IDENT. Apart from sloppiness in the annotation, this may also have been caused by the fact that the annotation tool registers such decisions only after the *apply* or *auto-apply* option has been selected.
3. Cases where multiple interpretations are possible.
4. Unclear guidelines. It was unclear whether titles and other leading material from news items should be considered part of the annotation task. It was unclear which appositions should be annotated with a PRED relation.

A more explicit formulation of the guidelines should eliminate most of the errors under 4. The fact that annotators must choose between IDENT and BRIDGE is a potential cause of disagreement that is probably harder to eliminate.

Visualization

The XML format of the MMAX annotation tool only supports viewing of the annotated material within the annotation tool itself. The possibilities for visualizing coreference information within this tool are somewhat limited, and furthermore, for users who only want to browse the annotation, installation of the tool is an undesirable overhead. We decided therefore, to convert the MMAX format into an XML format that can be inspected visually in a standard web-browser.[5]

We took the visualisation of coreference that was developed within the Norwegian Bredt project[6] as starting point. The actual visualisation is performed by a XSL stylesheet in combination with CSS and JavaScript. Documents are displayed as web- pages. All markables are bracketed. NPs that are part of some coreference relation appear in bold. The font color of anaphoric NPs indicates the nature of the coreference relation (i.e. IDENT, BRIDGE, ...). By moving the mouse over an NP, all NPs in the same coreference chain are highlighted. Different background colors indicate the relation of the other NPs to the selected NP (i.e. refers to or is referred to, direct or indirect reference). By clicking the left mouse button, all attributes of a markable are shown. An example is shown in figure 1.

3. Coreference resolution module

One of the major directions in the field of computational coreference resolution is the knowledge-based approach, in which there has been an evolution from the systems which require an extensive amount of linguistic and non-linguistic information (e.g. (Hobbs, 1978), (Rich and LuperFoy, 1988)) toward more knowledge-poor approaches (e.g. (Mitkov, 1998)).

In the last decade, machine learning approaches have become increasingly popular. Most of the machine learning approaches (e.g. (McCarthy and Lehnert, 1995), (Soon et al., 2001), (Ng and Cardie, 2002b), (Yang et al., 2003), (Ponzetto and Strube, 2006)) are supervised classification-based approaches and require a corpus annotated with coreferential links between NPs.

For the Dutch coreference resolution module we use a typical machine learning approach. We focus on identity relations. We start with detection of noun phrases in the documents after automatic preprocessing the raw text corpora. The following preprocessing steps are taken: rule-based tokenization using regular expressions. Dutch named entity recognition is performed by looking up the entities in lists of location names, person names, organization names and other miscellaneous named entities. We use a memory based part-of-speech tagger, text chunker and grammatical relation finder, each trained on the CGN corpus using the memory-based tagger-generator, MBT (Daelemans et al., 1996). Text chunking is splitting a sentence into noun and verb phrases. The grammatical relation finder detects relations between verb phrases and noun phrases in the text such as object, subject, or modifier relations.

On the basis of the preprocessed texts instances are created. We create an instance between every NP (candidate anaphor) and its preceding NPs (candidate antecedent), with a restriction of 20 sentences backwards. A pair of NPs that belongs to the same coreference chain gets a positive label; all other pairs get a negative label. For each pair of NPs a feature vector of 47 features is created containing information on the candidate anaphor, its candidate antecedent and the relation between both. The task of the classifier is to label each feature vector as describing a coreferential relation or not.

In a second step in this approach, a complete coreference chain has to be built between the pairs of NPs that were classified as being coreferential. We cluster overlapping pairs of NPs into groups and compute overlap between groups to determine the final coreference chains.

The feature vectors encode morphological-lexical, syntactic, semantic, string matching and positional information sources. The features can encode simple lexical information such as 'the anaphor is a definite noun or not' or positional information as 'distance in sentences' but also more complex information such as 'the anaphor and antecedent are synonyms' which requires a lookup in EuroWordNet(Vossen, 1998).

3.1. Cross validation

To evaluate the performance of the coreference resolution module, we run ten-fold cross validation experiments on 242 documents from the KNACK corpus. As our classifier we use the Timbl k nearest neighbor algorithm (Daelemans et al., 2004). We run experiments with a generational genetic algorithm(GA). Previous research (Daelemans et al., 2003) has shown that feature selection and algorithmic parameter optimization can lead to large fluctuations in the performance of a machine learning classifier. Genetic algorithms have been proposed as an useful method to find an optimal setting in the enormous search space of possible parameter and feature set combinations. We run experiments with a GA for feature set and algorithm parameter selection

[5] Unfortunately, highlighting does not work properly in Internet Explorer.

[6] `bredt.uib.no`

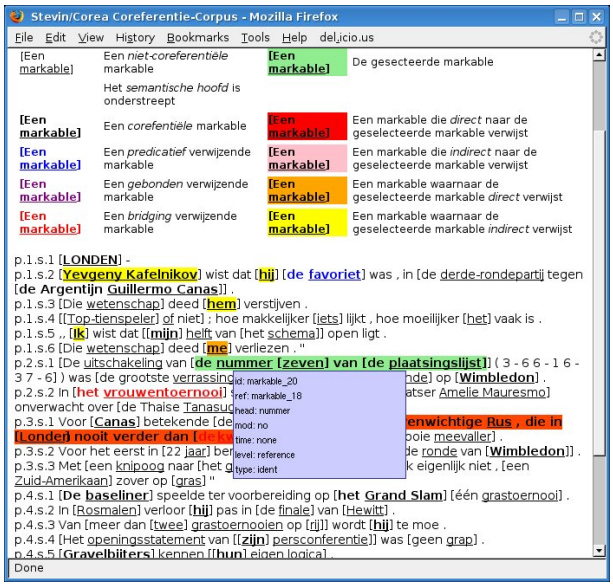

Figure 1: Screenshot of the visualization, with *de nummer zeven van de plaatsingslijst* (*the number 7 of the seeding*) selected.

MUC score	recall	precision	F-score
baseline	81.1	24.0	37.0
Timbl default	47.0	44.3	45.6
Timbl GA	36.8	70.2	48.2

Table 3: Micro-averaged F-score and accuracy computed in 10-fold c.v. experiments on 242 documents. Results of Timbl with default settings and with the settings as selected by the genetic algorithm.

of Timbl with 30 generations and a population size of 10. A detailed description of the genetic algorithm can be found in (Hoste, 2005).

We measure the MUC F-score on coreference chains as defined in Vilain et al. (1995). We also computed a baseline score by assigning each NP in the test set its most nearby NP as antecedent. The results are given in table 3. Timbl performs well above the baseline. Optimization with the GA leads to a higher precision for Timbl and overall higher F-score.

4. Extrinsic Evaluation

A more practical oriented evaluation is to test the usefulness of coreference relation information in an NLP application. We run experiments with an Information Extraction module for the medical domain, and measure the performance of this module with and without the coreference relation information predicted by our resolution module described in the previous section. We also present another application-oriented evaluation for the field of Question-Answering in which the effect of a simple rule-based coreference resolution module is measured.

4.1. Effect on Information Extraction

As a Information Extraction application we construct a Relation Finder which can predict medical semantic relations. This application is based on a version of the Spectrum medical encyclopedia (MedEnc) developed in the IMIX ROLAQUAD project, in which sentences and noun phrases are annotated with domain specific semantic tags (Lendvai, 2005). These semantic tags denote medical concepts or, at the sentence level, express relations between concepts. Example 6 shows two sentences from MedEnc annotated with semantic XML tags. Examples of the concept tags are *con_disease, con_person_feature* or *con_treatment*. Examples of the relation tags assigned to sentences are *rel_is_symptom_of* and *rel_treats*.

(6) <rel_is_symptom_of id="20"> Bij <con_disease id="2"> asfyxie</con_disease> ontstaat een toestand van <con_disease_symptom id="7"> bewustzijnverlies </con_disease_symptom> en <con_disease id="4">shock </con_disease> (nauwelijks waarneembare <con_person_feature id="8"> polsslag </con_person_feature> en <con_bodily_function id="13"> ademhaling </con_bodily_function>). </rel_is_symptom_of>
<rel_treats id="19"> Veel gevallen van <con_disease id="6"> asfyxie</con_disease> kunnen door <con_treatment id="14"> beademing </con_treatment>, of door opheffen van de passagestoornis (<con_treatment id="15"> tracheotomie </con_treatment>) weer herstellen. </rel_treats>

The core of the Relation Finder is a maximum entropy modeling algorithm trained on approximately 2000 annotated entries of MedEnc. Each entry is a description of a particular item such as a disease or body part in the encyclopedia and contains on average 10 sentences. It is tested on two separate test sets of 50 and 500 entries respectively. Our coreference module predicted coreference relations for the

noun phases in the data. We run two experiments with the Relation Finder, one using the predicted coreference relations as features, and one without these features.
The F-scores of the Relation Finder are presented in table 4 and show a modest positive effect for the experiments using the coreference information.

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

Table 4: F-Scores of Relation Finder

4.2. Effect on Question Answering

Mur (2006) describes a similar information extraction experiment, concentrating on relations where at least one of the arguments is a named entity, such as date-of-birth, age, capital-of, and founder-of. After adding coreference resolution, the number of extracted facts goes up with over 50% (from 93K to 145K). Incorporation of these facts into a Question Answering system leads to an improvement in accuracy of 5% (from 65% to 70%) on questions of the appropriate type. It should be noted, though, that Mur uses a simple rule-based coreference system, in combination with an automatically constructed knowledge base containing class labels for named entities. The accuracy of the newly added facts is therefore only 34%. Further improvements are probably possible by integrating the coreference resolution system described above.

5. Summary

We present the main outcomes of the Stevin COREA project: a corpus annotated with coreferential relations and the evaluation of the coreference resolution module developed in the project.
We discussed the corpus, the annotation guidelines, the annotation tool, and the inter-annotator agreement. We also showed a visualization of the annotated relations. We evaluated the coreference resolution module in two ways: with standard cross validation experiments to compare the predictions of the system to a hand-annotated gold standard test set, and a more practical oriented evaluation is to test the usefulness of coreference relation information in an NLP application.
The annotated data, the annotation guidelines, the visualization tools and web demo version of the coreference resolution application are available to all and will be distributed by the Dutch TST Centrale. [7]

Acknowledgments

The COREA project described in this paper was funded by the NWO-FWO STEVIN program.

6. References

G. Bouma, W. Daelemans, I. Hendrickx, V. Hoste, and A. Mineur. 2007. The corea-project, manual for the annotation of coreference in dutch texts. Technical report, University Groningen.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. Mbt: A memory-based part of speech tagger generator. In *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, pages 14–27.

W. Daelemans, V. Hoste, F. De Meulder, and B. Naudts. 2003. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, pages 84–95.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2004. TiMBL: Tilburg Memory Based Learner, version 5.1, reference manual. Technical Report ILK-0402, ILK, Tilburg University.

F. Fisher, S. Soderland, J. Mccarthy, F. Feng, and W. Lehnert. 1995. Description of the umass system as used for muc-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 127–140.

S. Harabagiu, R. Bunescu, and S. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, pages 55–62.

E. Hinrichs, S. Kübler, and K. Naumann. 2005. A unified representation for morphological, syntactic, semantic, and referential annotations. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 13–20.

L. Hirschman, P. Robinson, J. Burger, and M. Vilain. 1997. Automating coreference: The role of annotated training data. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

J.R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

V. Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

P. Lendvai. 2005. Conceptual taxonomy identification in medical documents. In *Proceedings of The Second International Workshop on Knowledge Discovery and Ontologies*, pages 31–38.

J. McCarthy and W. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, pages 1050–1055.

R. Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998/ACL-1998)*, pages 869–875.

MUC-7. 1998. Muc-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

J. Mur. 2006. Increasing the coverage of answer extraction by applying anaphora resolution. In *Fifth Slovenian and First International Language Technologies Conference (IS-LTC '06)*.

V. Ng and C. Cardie. 2002a. Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Proceedings of the 2002 Conference*

[7] www.tst.inl.nl/

on Empirical Methods in Natural Language Processing (EMNLP-2002), pages 55–62.

V. Ng and C. Cardie. 2002b. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

S. P. Ponzetto and M. Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 192–199.

E. Rich and S. LuperFoy. 1988. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 18–24.

W.M. Soon, H.T. Ng, and D.C.Y Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Y. Versley. 2006. Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-)reference. In *Proceedingsof Ambiguity in Anaphora ESSLLI Workshop*, pages 83–89.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

X. Yang, G. Zhou, S. Su, and C.L. Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41th Annual Meeting of the Association for Compuatiational Linguistics (ACL-03)*, pages 176–183.