Coreference Resolution for Extracting Answers
COREA

Gosse Bouma
Rijksuniversiteit Groningen

Walter Daelemans
Universiteit Antwerpen

March 7, 2005

1 Coreference Resolution for Extracting Answers: COREA

Coreference resolution is a key ingredient for the automatic interpretation of text. It has been studied mainly from a linguistic perspective, with an emphasis on establishing potential antecedents for pronouns. Practical applications, such as Information Extraction (IE), summarization and Question Answering (QA), require accurate identification of coreference relations between noun phrases in general. Computational systems for assigning such relations automatically, require the availability of a sufficient amount of annotated data for training and testing. For Dutch, annotated data is scarce and coreference resolution systems are lacking.

In this project, we aim to develop a robust system for assigning such relations automatically, and we will investigate the effect of making coreference relations explicit on the accuracy of systems for IE and QA. We will annotate a limited amount of application-specific corpus material, which is required for the evaluation of the coreference resolution system in the context of IE and QA.

The project contributes to the goals of Stevin by providing a robust coreference resolution system which is applicable in a range of applications for Dutch, such as information extraction, question answering and summerization. In addition, general guidelines for coreference annotation will become available and a tool will be developed to support the annotation of coreference in text. Finally, a limited amount of data annotated with coreferential information, including spoken language data, will be produced.

2 Description of the Proposed Research Project

2.1 Scientific aspects and innovative power

2.1.1 System development

The goal of the proposed project is to develop a robust system for the resolution of coreferential relations in text. Automatic, robust, domain independent, coreference resolution systems typically operate on top of or in combination with other NLP modules (such as a POS-tagger, a chunker, a named-entity recognizer or full syntactic analyzer), which provide *potential antecedents* for a given nominal phrase. The task of the resolution system is to select the most likely antecedent for the current nominal phrase. This decision is typically made on the basis of linguistic features of the potential antecedents, such as number and gender information, grammatical role, the distance between the current phrase and the potential antecedent, and ontological information, such as animacy. Note that using grammatical role as a source of information requires a shallow or full syntactic parser, and using ontological information requires a lexical resource such as WordNet.

Two different directions can be taken in research on computational coreference resolution: a knowledge-based approach and a corpus-based approach. Among the **knowledge-based approaches** to anaphora resolution, a distinction can be made between approaches, which generally depend upon linguistic knowledge (e.g. Hobbs (1978) and Lappin and Leass (1994)), and approaches in which discourse structure is taken into account (e.g. Grosz, Joshi and Weinstein (1995)). In all these approaches, there has been an evolution from systems requiring an extensive amount of linguistic and non-linguistic knowledge (e.g. Rich and LuperFoy (1988)) towards more knowledge-poor approaches (e.g. Mitkov (1998)). The systems depending on linguistic knowledge apply this lexical, morphological, syntactic and semantic knowledge through the use of constraints and preferences. Whereas the constraints are applied in order to remove bad antecedents, the preferences impose an ordering on the remaining candidate antecedents. Also discourse information has been used for automatic anaphora resolution. Especially centering (Grosz et al. 1995) and focusing theory (Sidner 1979) have been succesfully used.

Our proposed approach is corpus-based. **Corpus-based techniques** have become increasingly popular for the resolution of coreferential relations and was enabled by the creation of coreferentially annotated corpora such as MUC-6 and MUC-7. Dagan and Itai (1990), for example, derive collocation patterns from corpora and use these patterns to filter out unlikely antecedent candidates. Ge, Hale and Charniak (1998) use a statistical approach for the resolution of third person anaphoric pronouns. Machine learning techniques have gained popularity in the research on coreference resolution as well. Most machine learning approaches to coreference resolution are supervised techniques, such as the C4.5 decision tree learner (Quinlan 1993) as used by Aone and Bennett (1995), McCarthy (1996), Soon, Ng and Lim (2001) and others, maximum entropy learning as in Kehler (1997) and Luo, Ittycheriah, Jing, Kambhatla and Roukos (2004) and the Ripper rule learner (Cohen 1995) as in Ng and Cardie (2002). These approaches recast the problem as a classification task: a classifier is trained to decide whether a pair of NPs is coreferent or not. The pair of NPs is represented by a feature vector containing distance, morphological, lexical, syntactic and semantic information on the candidate anaphor, its candidate antecedent and also on the relation between both.

Not much research has been done yet on automatic Dutch coreference resolution. Furthermore, the existing research on this topic of op den Akker, Hospers, Lie, Kroezen and Nijholt (2002) and Bouma (2003)) falls within the knowledge-based resolution framework and only focusses on the resolution of pronominal anaphors. Bouma (2003), for example, presents an OT inspired anaphora resolution system for Dutch, which uses the Alpino-parser and Dutch EuroWordNet to determine the value of linguistic features such as number and gender, syntactic role, and animacy.

Automatic Dutch coreference resolution not restricted to resolution of the pronominal references, is still an unexplored research area. In this project, we plan to develop a resolution system based on machine learning techniques. In order to produce a successful coreference learner, we will address the following research questions:

- Since a set of well-defined features is crucial for the success of a given machine learner, we will thoroughly investigate the effect of different information sources. We will integrate shallow features (tags, chunks, named entities) with deep features, as provided by a full syntactic parser such as Alpino. The integration of deep and shallow processing in NLP is a topic that recently led to interesting results (e.g. in the *Deep Thought* EU project `http://www.project-deepthought.net/`). We will make use of EuroWordNet and other ontologies to investigate the contribution of semantic features containing information on synonyms, hypernyms, etc. We will also investigate alternatives to WordNet based on unsupervised learning from large unannotated corpora and the WWW. Finally, we will assess the informativity of the different types of information sources (morphological, lexical, syntactic, string-matching, semantic information) through the use of different feature selection techniques.

- Since coreference resolution data sets, as many other data sets, suffer from a severe imbalanced class distribution, in which one class (mostly the class to be learned) is largely underrepresented, we will investigate whether performance is hindered by the imbalanced class distribution in our data sets and we explore different strategies to cope with this skewedness. Previous studies (e.g. Lewis and Gale (1994) and Cardie and Howe (1997)) on other natural language data sets have shown that imbalanced data sets may result in poor performance of standard classification algorithms (e.g. decision tree learners, nearest neighbour and naive bayes methods). We will investigate different methods to cope with this skewedness, such as resizing training data sets or sampling, adjusting misclassification costs, learning from the minority class, adjusting the weighting of the examples, etc.

- Given the specifications for the resolution system (efficient, robust, reusable, accurate, unrestricted text as input, xml file indicating coreference chains as output), we will use the results of the machine learning experiments (including an empirical comparison of different machine learning methods) to develop a final design for the resolution system.

- The guiding criterion in the previous experiments will be high f-scores (high precision and recall). However, in deciding on the design of a robust and efficient reusable anaphora resolution system, a trade-off has to be found between efficiency and robustness on the one hand, and accuracy on the other. Some of the information sources studied are very expensive whereas others are cheaper (more shallow). At this stage we will perform optimization experiments using genetic algorithms (Daelemans 2003) to configure a design in terms of feature selection (choice of information sources) and algorithm parameters that constitutes a good trade-off between efficiency and accuracy (by combining both criteria in the fitness function of the genetic algorithm).

2.1.2 Corpus Development

Corpora annotated with coreferential information are a prerequisite for the development and evaluation of any resolution system. In the current project, we hope to gain access to such corpora by reuse of existing resources, and a limited amount of hand annotating new, application-oriented, material. As a side effect of this effort, general annotation guidelines for coreference annotation will become available, as well as tool for annotating coreference efficiently.

Reuse and Unification of Existing Resources

The most valuable resource for the development a Dutch coreference system is the corpus created recently at the University of Antwerp. A substantial Dutch corpus has been annotated with anaphorical relations between different types of noun phrases, including named entities, definite and indefinite NPs and pronouns. It is based on KNACK, a Flemish weekly magazine with articles on national and international current affairs. The corpus consists of 267 documents, in which 12,546 noun phrases are annotated with coreferential information.

In addition, we will consider using other general resources. The annotated corpus of op den Akker et al. (2002), for example consists of a number of texts from different types (newspaper articles, magazine articles and fragments from books) and contains 801 annotated pronouns. Bouma (2003) annotated a small corpus from the Volkskrant newspaper with 222 pronouns.

We will investigate to what extent these corpora can be converted into a common format, suitable for training and evaluation of an automatic coreference system, where the Antwerpen corpus is our point of reference.

Annotation Guidelines and Tools

The Antwerpen corpus was annotated on the basis of an explicit annotation scheme[1] which was based on the MUC-7 (MUC-7 1998) manual and the manual from Davies, Poesio, Bruneseaux and Romary (1998). It takes into account the critical remarks levelled against these annotation guidelines from

[1] `http://cnts.uia.ac.be/~hoste/manual_dutch.ps`

Kibble (2000) and van Deemter and Kibble (2000). These guidelines will also be the starting point for corpus development within the current project.

Annotation focusses primarily on coreference between noun phrases, as in (1-a) and (1-b), where both noun phrases refer to the same extra-linguistic entity.

(1) a. *Guy Verhofstadt*$_i$ had een onderhoud met *de premier van Nepal*$_j$. *De eerste minister van België*$_i$ heeft het met *de premier*$_i$ gehad over de wapenleveringen.
b. *Bert Anciaux*$_i$ zei dat *hij*$_i$ de aantijgingen beu was.

In addition, coreference in a modal context (as in (2-c)), bound anaphora (as in (2-a)), identity of sense anaphora (as in (2-b)) and time-dependent identities (as in (2-c)) are annotated. A TYPE attribute has been introduced to distinguish these forms of coreference from identity relations involving an extra-linguistic entity.

(2) a. Geen enkele Argentijn kan meer dan 1100 euro per maand van *zijn* rekening halen.
b. Enkele dagen eerder had de Waalse regering de voet op de instititutionele rem gezet om een einde te maken aan de tijdskredietpremies die de Vlaamse regering betaalt bovenop *die* van de federale overheid.
c. Chirac was in die tijd *voorzitter van de RPR en burgemeester van Parijs*. Medewerkers van *de president* beweren dat de terugkeer van Schuller *een politiek manoeuvre van links* is.

Other coreference relations, such as part-whole relations (as in (3-a), where the *motor* is part of *mijn eigen auto*), coreference between events and NPs as in (3-b), where *ramp* refers to *ten onder gaan*) and the relation between an antecedent and an inherent reflexive (3-c), are not annotated.

(3) a. "*De nieuwe motor* in mijn eigen auto was net op tijd klaar " , zei Verstappen .
b. De Estonia ging op 28 september ten onder in de Baltische Zee. Slechts 137 mensen overleefden *de ramp*.
c. Vrije Scholen worden net als andere gesubsidieerd door de overheid en hebben *zich* dan ook aan de onderwijsvoorschriften van die overheid te houden .

An example of the current annotation is given in figure 1.

For annotating new material, we will use one of the several tools which support coreference annotation. The Alembic system[2] has been used in the annotation of the Antwerpen Corpus. An alternative might be the Palenka system.[3] We will investigate to what extent the annotation task can be made easier by integrating a PoS tagger or syntactic parser. Access to PoS tags and information about constituent boundaries can be used to provide lists of potential antecedents. Choosing a suitable candidate from a list is typically much easier than identifying antecedents in raw text. In a later stage of the project, annotation will be supported by the automatic corefence resolution system. In that case, the system will suggest an antecedent, and the annotator only has to correct errors. Again, this typically speeds up the annotation task, even if the resolution system itself can still be improved.

Application Oriented Corpora

Our aim is to develop a coreference resolution system whose performance is robust and accurate enough that it can be used in applications. To evaluate this, we will also annotate a limited amount of

[2] `http://www.mitre.org/tech/alembic-workbench`

[3] `http://pers-www.wlv.ac.uk/~le1825/`

```
<COREF ID="1528" MIN="conflict">Het conflict over het
 grensgebied</COREF> is zo oud als <COREF ID="1464"><COREF
 ID="1451">India</COREF> en <COREF ID="1459">Pakistan</COREF>
</COREF>.

 Er zijn <COREF ID="1454">twee Kashmirs</COREF>. <COREF ID="1455"
 TYPE="IDENT" REF="1454">De Indiase, vrijwel autonome deelstaat
 Jammu&ndash;Kashmir en het Pakistaanse Azad Kashmir, Vrij
 Kashmir</COREF>. In de praktijk is er van autonomie of vrijheid voor
 <COREF ID="1456" TYPE="IDENT" REF="1454">de beide Kashmirs</COREF>
 geen sprake, want <COREF ID="1457" TYPE="IDENT" REF="1454">ze</COREF>
 zijn sinds jaar en dag <COREF ID="1458" TYPE="IDENT" REF="1454"
 MIN="twistappel">d&eacute; twistappel tussen <COREF ID="1466"
 TYPE="IDENT" REF="1464"><COREF ID="1460" TYPE="IDENT"
 REF="1459">Pakistan</COREF> en <COREF ID="1463" TYPE="IDENT"
 REF="1451">India</COREF></COREF></COREF>.
```

Figure 1: Sample of coreference annotation based on the MUC-annotation scheme.

application specific data. In particular, we will develop a corpus which is representative for resolving coreference in dialogue, and a corpus which is representative for IE and QA tasks.

- A suitable portion of the recently completed Corpus of Spoken Dutch (CGN) will be enriched with coreference relations. CGN contains dialogue fragments which pose the same problems for coreference resolution as those encountered in spoken dialogue systems. The material in CGN has already been annotated with PoS tags and syntactic relations. This information can be used to boost the annotation process, as described in the previous section.

- A fragment will be annotated which is typical for (domain specific) IE and QA tasks. We will annotate a suitable fragment of medical text. Through the NWO IMIX programme, we have access to a medical encyclopedia, and substantial collections of medical text from various internet sources. Part of this material has already been annotated at Tilburg University with domain concepts. This ontological knowledge can be used to facilitate annotation.

2.1.3 Application Oriented Evaluation

A coreference resolution system is usually evaluated by computing what the accuracy of automatically assigned coreference relations is on a representative but unseen (in training and development) portion of the annotated corpus. In this project we will perform such an internal evaluation, based on cross-validation of the data in the machine learning experiments.

In addition, we will investigate what the effect of coreference resolution is on a typical IE or QA task. Coreference resolution is essential for extracting information about persons, organizations, events, *etc.* from running text. An example is given in figure 2, where various linguistic forms are used to refer to the same person (*Mugabe*). Extracting the information that is provided about Mugabe from this text requires that the coreference of the highlighted linguistic expressions can be established. Figure 3 gives two examples from the CLEF 2003 QA track, where establishing coreference is crucial for determining that the text fragment provides an answer to the question.

Noch de luide veroordeling door de Britse regering noch de ’stille diplomatie’ van Zuid-Afrika konden **Mugabe** overtuigen zich beter te gedragen. Tijdens een serie topontmoetingen en conferenties beloofde **hij zijn** eigen wetten te respecteren en vrije en eerlijke verkiezingen toe te staan. (...) De vraag is: zal **Mugabe** toch verliezen? Aan de ene kant zouden de Zimbabwanen, na 22 jaar van wanbeleid, maar al te graag van **hem** af zijn. Aan de andere kant, speelt **de oude man** nu al vals. (...) Onverschrokken moedigde **de president** de oorlogsveteranen aan de boerderijen van blanken te bezetten.

Figure 2: Illustration of the relevance of coreference resolution for information extraction systems.

- *Welke kracht had de aardbeving waardoor het noorden van Japan werd getroffen ?*
- Een snelweg in Bekkai op Hokkaido, het meest noordelijke eiland van Japan, scheurde als gevolg van **de krachtige aardbeving in de Stille Oceaan**. **De onderzeese aardbeving** was de zwaarste die de regio in een kwart eeuw trof en had een kracht van 7,9 op de schaal van Richter.
- *Hoe heet de zoon van Kim Il Sung ?*
- Het lichaam van de Noordkoreaanse leider **Kim Il Sung** is zaterdag op zijn eerste sterfdag in het openbaar tentoongesteld in de hoofdstad Pyongyang. De plechtigheid was er volgens waarnemers op gericht te laten zien dat **Kims** zoon en opvolger Kim Jong Il daadwerkelijk aan de macht is.

Figure 3: CLEF 2003 questions and text snippets containing an answer. Resolving coreference relations is essential for establishing that the text contains an answer to the question.

For the TREC QA data sets, Morton (2000) and Watson, Preiss and Briscoe (2003) have shown that adding coreference resolution to a QA system may have a positive effect on the accuracy of the overall system. Watson *et al.* point out that coreference resolution in scientific text may be harder than in the newspaper text used for the TREC QA competition, as scientific text tends to be more complex and contains relatively high proportions of definite descriptions, which are the most challenging to resolve.

Within the NWO IMIX project, Groningen University has started to develop an open domain QA system for Dutch, which will participate in the upcoming QA tracks of the annual CLEF conference. The current system relies heavily on syntactic information, both for direct question answering as well as for various (off-line) information extraction tasks which may help QA. Based on the idea that all TREC and CLEF data sets released to date contain questions asking for specific relations (such as *country-capital, country-leader, country-currency, abbreviation-full term, event-location, event-year, etc.*), a system has been developed which searches the text collection exhaustively for such relations. To this end, the full text collection has been parsed by the Alpino system (van der Beek, Bouma and van Noord 2002), and the resulting dependency trees are stored as XML. Jijkoun, Mur and de Rijke (2004) demonstrate that the patterns which extract specific relations from dependency representations are considerable more effective than systems which use regular expressions to search raw text. The Amsterdam submission for QA@CLEF 2004 (Jijkoun, Mishne, de Rijke, Schlobach, Ahn and Müller 2004) used the same technique for their Dutch QA system. The dependency trees for the Dutch corpus were produced by the Alpino system. Although Amsterdam provided the only submission for the Dutch monolingual task, it should be noted that it achieved a far higher score than systems for other European languages on a similar task.

In this project, we will evaluate how the recall of IE can be improved by adding coreference information. We will investigate the effect of adding coreference to the IE tasks necesarry for (open-domain) QA. In addition, we will investigate how an IE task in the medical domain may profit from coreference information.

Language and Computing (L&C) will design the Information Extraction and QA scenarios for the medical domain that will be used in this evaluation. These scenarios will vary in complexity and will be designed by analogy with real cases from companies and organizations, active in healthcare or the pharmaceutical sector, L&C was confronted with. The IE system will be evaluated on its ability to fill the IE templates and the correctness of the QA results will be verified. The final evaluation and reporting of the IE and QA systems will be performed by L&C's medical linguists and ontologists. An evaluation document will be delivered for both the systems, with comments and suggestions for the improvement of the underlying technologies and hence for the maturation of the IE and QA systems

2.2 Economic aspects

Boosting the Performance of Intelligent Text Processing

There is a rapidly increasing market for tools that have the ability to search intelligently for information in unrestricted text and that help in the mining and management of the knowledge implicit in unrestricted text. Such tools are part of systems that perform automatic summarization, information extraction from unrestricted text, and question answering. Coreference resolution is a key technique for making such systems more powerful and is essential for almost every conceivable form of semantic processing of text. There have recently been a series of workshops which stress the importance of coreference resolution for applications (see section 8 below). It has been shown that resolving coreference relations can improve the recall of information extraction from free text and can also have a positive effect on the performance of QA systems.

By developing an automatic coreference resolution system for Dutch, we make this technology available for intelligent information processing systems which have to deal with Dutch text and spoken language. We aim to build a coreference resolution system based on machine learning technology which is reusable in a wide range of applications, such as information extraction, question answering and summerization. By developing and evaluating our system in the context of realistic applications, we will ensure that the resulting system can be used to obtain real performance improvements.

The presence of Language and Computing will ensure realistic application scenarios and thorough feasibility study of application of the results in medical text mining.

Resources for Coreference Research

The project also develops an infrastructure for annotating text with coreference relations interactively. We will provide general annotation guidelines, a tool for annotating text with coreference relations, as well as a limited amount of corpus data. This infrastructure will contribute to future research on coreference. The corpus material can be used to develop and evaluate alternative coreference systems. The guidelines and annotation tool can be used in future projects which seek to annotate more substantial corpora (as mentioned in the Stevin programme priorities), or in which coreference annotation is combined with other layers of annotation.

2.3 Contribution to the Stevin programme

The proposed project contributes to Stevin priorities by

- Designing and evaluating a coreference resolution system for Dutch, applicable in a wide range of intelligent text processing applications, such as Information Extraction and Question Answering,

- Development of an infrastructure for coreference research (consisting of annotation guidelines, an annotation tool, and a limited amount of annotated data) which may be used for evaluating alternative systems or for annotating additional material,

- Enriching part of the Corpus of Spoken Dutch with coreference relations.

  More specifically, the proposed project will contribute to the following language technology priorities mentioned in the call for proposals:

  - Richly annotated monolingual Dutch corpora,
  - Semantic analysis,
  - Monolingual information extraction,
  - QA solutions.

  With these priorities, the proposed project addresses the three main aspects of digital language infrastructure development for Dutch: resources, strategic research, and to a lesser extent applications. The project would also integrate the complementary expertise of the Antwerp (machine learning for NLP) and Groningen (full parsing for deep NLP) groups and contributes to knowledge transfer by having a company (Language and Computing) define realistic application scenarios and do the evaluation of the developed system.

2.4 IPR and standards

All deliverables of the project (annotation guidelines, annotation tool, corpora, coreference resolution tool, results of evaluation) will be made publicly available through the TST-Centrale. In particular, we will ensure that there are no copy-right or other restrictions on the text material selected for annotation. We expect to use tools available in the public domain already for research and education and newly developed software will become available under the same conditions.

Annotation guidelines will broadly follow the annotation conventions set for English (i.e. in MUC), but may improve on these conventions where we feel this is necessary. Corpora will be made available as XML documents, and will follow standards for encoding of text corpora, to the extent that these apply to the current annotation effort.

Duplication is not an issue as the scarce already existing resources (mainly annotated corpora) will be integrated into this project.

2.5 Co-ordination and project management

The project will be managed by the two principal investigators.

Regular meetings (at least 4 per year) will be scheduled between the two university research groups and the industrial partner, to ensure that development and annotation efforts are in sync, and to collaborate on application oriented evaluation.

Each work package will be coordinated by one of the academic partners who will be responsible for producing the associated deliverables and organising the work. Standard best practice methods (web-site, groupware, cvs, bugzilla, etc.) will be used for joint development and communication between the partners, and for dissemination of the results.

2.6 Evaluation, validation and success criteria

The project will deliver:

- Coreference Annotation Guidelines,
- An Annotation Tool,
- Annotated Data,
- A robust and reusable automatic coreference resolution system,
- Results of evaluating the system on a general coreference resolution task (including on spoken language data),
- Results of evaluating the system on a QA and IE application

The quality of the annotation guidelines and annotated data will be established by performing an inter-annotator agreement experiment. By annotating a text fragment twice (i.e. by annotators of the two university groups) and measuring the amount of agreement, we learn whether the guidelines are sufficiently clear and to what extent the annotation process itself is prone to errors. The experiment will be conducted at the beginning of the project (so guidelines and procedures can be adapted) and towards the end of the project (so a figure can be given for the expected percentage of remaining errors in annotated text).

The coreference system itself is rigorously evaluated internally by standard cross-validation methods. Evaluation will also be performed application-based on an IE and a QA task. Performance will be compared with results reported for other systems (on other languages and corpora).

The project is successfull if it delivers

- an infrastructure for coreference research which will be useful and which will be used in future research and corpus development involving coreference,
- a coreference resolution tool which can be shown to have a positive effect on intelligent information processing tasks, such as IE and QA.

3 Work Programme

Workpackages

Work-package coordinating partner is in boldface.

WP0 Management

Tasks: Setting up and maintaining infrastructure for cooperation, communication, reporting and dissemination of results. (groupware, cvs, website) and dissemination of results. Organizing trimestrial meetings.
Responsible: **UA, RUG**

WP1 Guidelines

Tasks: Development of annotation guidelines starting from the CNTS (UA) guidelines. Defining XML DTD for annotation. Measuring inter-annotator agreement.
Responsible: **UA**, RUG
Deliverable: Annotation Guidelines (report). [month 2]
Risks and alternatives: Validation of the guidelines may indicate low inter-annotator agreement. In that case the protocol may be adapted by leaving out some coreference types.

WP2 Annotation Tool

Tasks: An existing annotation tool will be adapted to the task at hand. Annotated text will be stored as XML, following a DTD which defines an XML syntax for the annotation guidelines. The tool should support easy editing, ensure that annotated material is conformant with the DTD, and should allow for visualisation of output other NLP modules, such as a PoS tagger, parser, and integration of output from an (experimental) coreference resolution system.
Responsible: UA, **RUG**.
Dependencies: Depends on WP1.
Deliverable: Annotation Tool (software). [month 2]
Risks and alternatives: Given the experience in Antwerp with Alembic, no problems are foreseen.

WP3 Corpus preparation and annotation

Tasks: Existing corpus material (mainly KNACK) will be converted to the new annotation format. New material will be annotated using the annotation tool according to the guidelines by student-assistents: CGN spoken language material, and IMIX medical text.
Responsible: UA, **RUG**.

Dependencies: Depends on WP1 and WP2.
Deliverable: Annotated corpora (resources). [month 4]
Risks and alternatives: in case other corpus annotation efforts are funded within Stevin in parallel, close cooperation with this project in terms of guidelines, annotation tool and corpus selection will be pursued.

WP4 Coreference Resolution Tool This workpackage consists of four subsets of tasks:

- WP 4.1 Feature Engineering. Selection and preparation of tools (available with the partners) for the construction of features for the machine learning system: shallow features (lemma, part of speech, chunk, named entity, grammatical relation); syntactic features (ALPINO); semantic features using EuroWordnet Dutch; semantic features using unsupervised learning from unannotated corpora and using information extracted from WWW.
  Responsible: **UA**, RUG
  Dependencies: Depends on partial availability of WP3 results
  Deliverables:
  Report on feature engineering (publication). [month 8]
  Report on selected tools for feature construction. [month 8]
  Scripts and tools for feature construction from NLP tool output (software). [month 8].

- WP 4.2 Machine Learning Experiments. Feature selection experiments. Solutions for skewed data problem. Comparative Machine Learning experiments. Optimization of feature selection and algorithm parameter settings using genetic algorithms.
  Responsible: **UA**, RUG
  Dependencies: Depends on WP4.1 results
  Deliverables: Report on feature selection and ML algorithm optimization (publication). [month 16]

- WP4.3 Validation. Internal validation by cross-validation. Evaluation of the optimised system on a held-out dataset. The coreference resolution tool will be evaluated on representative corpus samples, unseen in the development and training of the system. The result is an accuracy figure for coreference resolution. The results of error analysis will be reported, and the accuracy will be compared to that of comparable systems for other languages.
  Responsible: **UA**, RUG
  Dependencies: Depends on WP4.2 results
  Deliverables: Report on final experiments (publication). [month 17]

- WP4.4 System development. Integration of feature construction software, NLP tools, ML software and optimised settings into a reusable, server-based tool. Documentation of the tool. The system is to operate on unrestricted text, enriched with linguistic information by other modules (such as a PoS tagger, named entity tagger, syntactic parser, domain specific ontological knowledge), and produces output conformant with the XML DTD.
  Responsible: UA, **RUG**
  Dependencies: Depends on WP4.2 and WP4.1 results
  Deliverables:
  Robust coreference resolution system with documentation (software, report, publication).

[month 18]

Risks and alternatives: In case of insufficient accuracy of the developed system for use in application-dependent evaluation, gold standard material will be used to investigate whether coreference resolution helps in IE and QA.

WP5 Application-dependent evaluation This workpackage evaluates the developed coreference resolution system in the context of real applications, the scenarios of which are developed by the commercial partner, Language and Computing.

- WP5.1 QA Application. Integration of coreference tool into the open domain QA system developed at RUG. Effect of coreference resolution on recall and precision of the overall application will be measured and output will be evaluated. Responsible: L&C, UA, **RUG**
  Dependencies: Depends on WP4 results
  Deliverables: Report on evaluation (publication). [month 24]

- WP5.2 IE Application. Integration of coreference tool into an IE application for factoid extraction in medical domain. Effect of coreference resolution on recall and precision of the overall application will be measured and output will be evaluated. Responsible: L&C, **UA**, RUG
  Dependencies: Depends on WP4 results
  Deliverables: Report on evaluation (publication). [month 24]

4 International Perspective

Research on coreference resolution, as reviewed in section 6.1, has focussed primarily on the development of automatic coreference resolution systems that were evaluated on hand annotated corpora. For this type of research, the MUC conferences were very influential. Recently, there has been a growing interest in tuning and evaluating coreference resolution in the context of realistic applications. Several recent workshops were dedicated to this topic (i.e. the ACL 2004 workshop on *Reference Resolution and its Applications*,[4] the EACL 2003 workshop on *the computational treatment of anaphora*,[5] a series of *Discourse Anaphora and Anaphor Resolution Colloquia* (hosted by the University of Lisbon in 2004[6] and the *2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization* (Venice)[7]). The current project contributes to this line of research by explicitly addressing the performance of automatic coreference resolution in intelligent text processing applications.

The project will combine state-of-the-art machine learning techniques with various linguistic analysis modules developed by the project partners.

The work on robust, deep, syntactic parsing at Groningen University has been recognized internationally as an important contribution to the field of statistical parsing, witness many recent presentations. Recently, Groningen University has started to develop an open domain question answering

[4] `http://acl.eldoc.ub.rug.nl/mirror/acl2004/refres/index.html`

[5] `http://www.itri.bton.ac.uk/˜Kees.van.Deemter/anaphora-program.html`

[6] `http://daarc2004.di.fc.ul.pt`

[7] `http://sisley.cgm.unive.it/ARQAS/Symposium2003.html`

system for Dutch, which is innovative in that it is based on full parsing of the text collections (consisting of 75 million words in the case of CLEF) used for QA. The results of this research have already been integrated in the Dutch QA system of the University of Amsterdam.

The computational linguistics group in Groningen cooperates with Stanford University, Ohio University, and the University of Tübingen. An important theme in these cooperations is the development of robust and efficient parsers for linguistically motivated grammar formalisms, topics which play a role in the proposed project.

The language technology group in Antwerp is internationationally acknowledged as an expertise center for the application of machine learning techniques to natural language processing. This work has, apart from theoretical results, led to the development of a suite of natural language processing tools in many languages, including Dutch (morphological analyzers, text to speech, chunkers, relation finders, named-entity recognition, word sense disambiguation, etc. Publications, projects and other results of this work can be found on the center's website: `cnts.uia.ac.be`. Recently, the language technology group has started focusing on text mining applications including information extraction, summarization, and question answering. CNTS has close international research contacts and cooperations with among others Bar Ilan University (Ido Dagan), University of Manchester (Text Mining for Bio-informatics), University of Vienna (Artificial Intelligence), University of Geneva (Text Mining), ILSP Athens, BBC, and SYSTRAN (summarization) on the topics of machine learning of language and information extraction applications, topics which play a role in the proposed project.

Language and Computing NV (L&C, `http://www.landcglobal.com/`) was founded in 1998. Today L&C consists of a staff of PhDs with extra degrees in computer science or computer linguistics, a group of doctors with extra experience in IT and an experienced sales- and marketing team. L&C is headquartered in Belgium, with a Sales Office in Virginia, US. The company specializes in information management technology for the healthcare and pharmaceutical markets. These solutions are a middleware platform for Natural Language Understanding (NLU) technology to capture the real meaning of free text documents and use this knowledge to make the information processable by computers. L&C offers solutions for semantic indexing of free text documents, information retrieval and extraction, terminology management and automated clinical coding and also company-tailored solutions. L&C has customers world-wide among all the players in the healthcare domain, e.g., hospitals, pharmaceutical companies and software developers.

5 Literature

Selection of Publications

1. Bouma, G., Finite state methods for hyphenation, *Journal of Natural Language Engineering*, **9**, 2001, pp. 5–20.

2. van der Beek, L., Bouma, G. and van Noord, G., Een brede computationele grammatica voor het Nederlands, *Nederlandse Taalkunde*, **7**(4), 2002, pp. 353–374.

3. Bouma, G., Malouf, R. and Sag, I., Satisfying constraints on adjunction and extraction, *Natural Language and Linguistic Theory*, **19**, 2001, pp. 1–65. rapport in opdracht van de Nederlandse Taalunie.

4. van Noord, G., Bouma, G., Koeling, R. and Nederhof, M.-J., Robust grammatical analysis for spoken dialogue systems, *Journal of Natural Language Engineering*, **5**, 1999, pp. 45–93.

5. Bouma, G. and Schuurman, I., De positie van het Nederlands in Taal- en Spraaktechnologie, 1998.

6. Buchholz, S. and Daelemans, W., Complex answers: a case study using a WWW question answering system. *Natural Language Engineering*, **7**(4), 2001, pp. 301–323..

7. Daelemans, W., van den Bosch, A. and Zavrel, J., Forgetting exceptions is harmful in language learning, *Machine Learning*, **34**(1/3), 1999, pp. 11-43.

8. Daelemans, W., Memory-Based Language Processing, *Journal of Experimental and Theoretical AI* (JETAI), **11**(3), 1999.

9. van Halteren, H., Zavrel, J. and Daelemans, W., Improving accuracy in word class tagging through combination of machine learning systems, *Computational Linguistics*, **27**(2), 2001, pp. 199–230.

10. Hoste, V., Hendrickx, I., Daelemans, W. and van den Bosch, A., Parameter Optimization for Machine-Learning of Word Sense Disambiguation, *Natural Language Engineering*, Special Issue on Word Sense Disambiguation Systems, **8**(4), 2002, pp. 311–325.

International Literature

1. Aone, C. and Bennett, S.W., Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL95)*, 1995, pp. 122–129.

2. Grosz, B., Joshi, A. and Weinstein, S., Centering: a framework for modeling the local coherence of discourse, *Computational Linguistics*, **21**(2), 1995, pp. 203–225.

3. Hirst, G., Anaphora in Natural Language Understanding: A Survey, *Lecture Notes in Computer Science*, **119**, 1981.

4. Hobbs, J., Resolving pronoun references, *Lingua*, **44**, 1978, pp. 311–338.

5. Lappin, S. and Leass, H., An algorithm for pronominal anaphora resolution, *Computational Linguistics*, **20**(4), 1994, pp. 535–561.

6. McCarthy, J., A trainable approach to coreference resolution for information extraction, *PhD thesis*, 1996.

7. Mitkov, R., *Anaphora Resolution*, 2002.

8. Ng, V. and Cardie, C., Improving machine learning approaches to coreference resolution, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, 2002, pp. 104–111.

9. Poesio, M. and Vieira, R., A Corpus-Based Investigation of Definite Description Use, *Computational Linguistics*, **24**(2), 1998, pp. 183–216.

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

References

Aone, C. and Bennett, S., Evaluating automated and manual acquisition of anaphora resolution strategies, *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL95)*, 1995, pp. 122–129.

Bouma, Gerlof, Doing dutch pronouns automatically in optimality theory, *Proceedings of the EACL 2003 Workshop on The Computational Treatment of Anaphora*, 2003.

Cardie, C. and Howe, N., Improving minority class prediction using case-specific feature weights, *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, 1997, pp. 57–65.

Cohen, W. W., Fast effective rule induction, *in* A. Prieditis and S. Russell (eds), *Proceedings of the 12th International Conference on Machine Learning*, Morgan Kaufmann, Tahoe City, CA, 1995, pp. 115–123.

Daelemans, W., Hoste, V., De Meulder, F. and Naudts, B., Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language, *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, Cavtat-Dubrovnik, Croatia, 2003, pp. 84–95.

Dagan, I. and Itai, A., Automatic processing of large corpora for the resolution of anaphora references, *Proceedings of the 13th International Conference on Computational Linguistics*, pp. Vol. III, 1990, 1–3.

Davies, S., Poesio, M., Bruneseaux, F. and Romary, L., Annotating coreference in dialogues: Proposal for a scheme for mate, 1998.

Ge, N., Hale, J. and Charniak, E., A statistical approach to anaphora resolution, *Proceedings of the Sixth Workshop on very Large Corpora*, 1998, pp. 161–170.

Grosz, B., Joshi, A. and Weinstein, S., Centering: a framework for modeling the local coherence of discourse, *Computational Linguistics* **21**(2), 1995, pp. 203–225.

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

Jijkoun, V., Mishne, G., de Rijke, M., Schlobach, S., Ahn, D. and Müller, K., The University of Amsterdam at QA@CLEF 2004, *CLEF 2004 Working Notes*, Bath, 2004.
*http://clef.isti.cnr.it/2004/working_notes/CLEF2004WN-Co ntents.html

Jijkoun, V., Mur, J. and de Rijke, M., Information extraction for question answering: Improving recall through syntactic patterns, *Coling 2004*, Geneva, 2004, pp. 1284–1290.

Kehler, A., Probabilistic coreference in information extraction, *in* R. I. Providence (ed.), *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP97)*, 1997.

Kibble, R., Coreference annotation: Whither?, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, 2000, pp. 1281–1286.

Lappin, S. and Leass, H., An algorithm for pronominal anaphora resolution, *Computational Linguistics* **20**(4), 1994, pp. 535–561.

Lewis, D. and Gale, W., Training text classifiers by uncertainty sampling, *Proceedings of the Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 3–12.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N. and Roukos, S., A mention-synchronous coreference resolution algorithm based on the bell tree, *in* S. Barcelona (ed.), *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL04)*, 2004, pp. 136–143.

McCarthy, J., *A Trainable Approach to Coreference Resolution for Information Extraction*, PhD thesis, Department of Computer Science, University of Massachusetts, Amherst MA, 1996.

Mitkov, R., Robust pronoun resolution with limited knowledge, *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98)*, Montreal, Canada, 1998, pp. 869–875.

Morton, T., Coreference for NLP applications, *Proceedings of the 38 annual meeting of the ACL*, Hong Kong, 2000, pp. 173–180.

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

Ng, V. and Cardie, C., Combining sample selection and error-driven pruning for machine learning of coreference rules, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, 2002, pp. 55–62.

op den Akker, H., Hospers, M., Lie, D., Kroezen, E. and Nijholt, A., A rule-based reference resolution method for dutch discourse, *Proceedings 2002 Symposium on Reference Resolution in Natural Language Processing*, 2002, pp. 59–66.

Quinlan, J., *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA, 1993.

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

Sidner, C., *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*, PhD thesis, Massachusetts Institute of Technology, 1979.

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

van Deemter, K. and Kibble, R., On coreferring: Coreference in muc and related annotation schemes, *Computational Linguistics* **26**(4), 2000, pp. 629–637.

van der Beek, L., Bouma, G. and van Noord, G., Een brede computationele grammatica voor het Nederlands, *Nederlandse Taalkunde* **7**(4), 2002, pp. 353–374.

Watson, R., Preiss, J. and Briscoe, T., The contribution of domain-independent robust pronominal anaphora resolution to open-domain question answering, *International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, Venice, 2003.
*http://sisley.cgm.unive.it/ARQAS/papers.html