Learning Dutch Coreference Resolution

Véronique Hoste and Walter Daelemans

CNTS-language Technology Group, University of Antwerp

Abstract

This paper presents a machine learning approach to the resolution of coreferential relations between nominal constituents in Dutch. It is the first significant automatic approach to the resolution of coreferential relations between nominal constituents for this language. The corpus-based strategy was enabled by the annotation of a substantial corpus (ca. 12,500 noun phrases) of Dutch news magazine text with coreferential links for pronominal, proper noun and common noun coreferences. Based on the hypothesis that different types of information sources contribute to a correct resolution of different types of coreferential links, we propose a modular approach in which a separate module is trained per NP type.

1 The task of coreference resolution

Although largely unexplored for Dutch, automatic coreference[1] resolution is a research area which is becoming increasingly popular in natural language processing (NLP) research. It is a weakness and therefore a key task in applications such as machine translation, automatic summarization and information extraction for which text understanding is of crucial importance.

But the resolution of coreferential relations is a complex task since it requires finding the correct antecedent among many possibilities. Furthermore, as shown in example (1) below, it involves different types of knowledge: morphological and lexical knowledge such as number agreement and knowledge about the type of noun phrase, syntactic knowledge such as information about the syntactic function of anaphor and antecedent, semantic knowledge which allows us to recognize synonyms and hyperonyms or which allows distinctions to be made between person, organization or location names, discourse knowledge, world knowledge, etc.

(1) Op 9 november 1983 werd **Alfred Heineken** samen met **zijn** chauffeur ontvoerd. **De kidnappers** vroegen 43 miljoen gulden losgeld. Een bescheiden bedrag, vonden **ze** zelf.
English: On 9 November 1983 **Alfred Heineken** and **his** driver were kidnapped. **The kidnappers** asked a ransom of 43 million guilders. A modest sum, **they** thought.

Whereas corpus-based techniques have become the norm for many other natural language processing tasks (such as part-of-speech tagging, parsing, grapheme-to-phoneme conversion, etc.), the field of computational coreference resolution is still highly knowledge-based, also for Dutch. Among these **knowledge-based approaches** to coreference resolution, a distinction can be made be-

[1] The discussion whether a given referring link between two nominal constituents can be qualified as coreferential, anaphoric or not is beyond the scope of this paper. We will use both terms interchangeably as is also done in most of the work on computational coreference resolution.

tween approaches which generally depend upon linguistic knowledge (Lappin and Leass 1994, Baldwin 1997), and the discourse-oriented approaches, in which discourse structure is taken into account, as in Grosz, Joshi and Weinstein (1995). Beside the fact that not much research has been done yet on automatic coreference resolution for Dutch, the existing research on this topic from op den Akker, Hospers, Lie, Kroezen and Nijholt (2002) and Bouma (2003) falls within the knowledge-based resolution framework and focuses on the resolution of pronominal anaphors. In this paper, we take another perspective and present a machine learning approach to the resolution of coreferential relations between different types of nominal constituents. It is the first corpus-based resolution approach proposed for Dutch. The corpus-based strategy was enabled by the annotation of a new corpus with coreferential relations between noun phrases.

The remainder of this paper is structured as follows. In the following section, we briefly describe the construction and the annotation of the KNACK-2002 corpus. In section 3, we continue with a description of the construction of the data sets. More specifically, we look at the different preprocessing steps that were taken, we consider the construction of positive and negative instances for the training data and test instances for the test data and we motivate the use of three smaller data sets (one for each NP type) instead of one single data set for training and testing. Section 4 gives an overview of the different features that were incorporated in the feature vectors for the machine learning methods we are using. In section 5, we introduce the two machine learning methods which are used for the experiments: memory-based learning and rule-induction. We continue with a description of the experimental setup, viz. the two-step learning approach and the evaluation methodology. Section 6 gives an overview of the experimental results in comparison to two baseline scores. We end this section with a qualitative error analysis of three KNACK-2002 documents. We conclude with a summary.

2 KNACK-2002

Lacking a substantial Dutch corpus of coreferential relations between different types of noun phrases, including named entities, definite and indefinite NPs and pronouns, we annotated a corpus ourselves. This annotation effort was crucial since the existing corpora for Dutch only contain coreferential relations for pronouns and are rather small. The annotated corpus of op den Akker et al. (2002), for example, consists of different types of texts (newspaper articles, magazine articles and fragments from books) and contains 801 annotated pronouns. Another corpus for Dutch was annotated by Bouma (2003). It is based on the Volkskrant newspaper and contains coreferential relations for 222 pronouns.

Our Dutch coreferentially annotated corpus is based on KNACK, a Flemish weekly news magazine with articles on national and international current affairs. KNACK covers a wide variety of topics in economical, political, scientific, cultural and social news. For the construction of this Dutch corpus, we used a selection of articles of different lengths from KNACK, which all appeared in the first ten weeks of 2002.

For the annotation of the Dutch news magazine texts, the following strategy was taken. First, an annotation scheme was developed containing a set of guidelines for marking up coreferences between noun phrases. On the basis of this annotation scheme, all texts were annotated by two annotators from a pool of five native speakers with a background in linguistics. After the individual coreference annotation by both annotators, they verified all annotations together in order to reach a single consensus annotation rather than keeping several, possibly differing, annotations. In case of no agreement, the relation was not marked. This decision was based on the observations of Hirschman, Robinson, Burger and Vilain (1997) that more than half (56%) of the errors were missing annotations and that 28% of the errors represented "easy" errors (such as the failure to mark headlines or predicating expressions).

The annotation scheme[2] for our Dutch corpus was based on the existing annotation schemes for English. We took the MUC-7 (MUC-7 1998) manual and the manual from Davies, Poesio, Bruneseaux and Romary (1998) as source and we also took into account the critical remarks on these schemes by van Deemter and Kibble (2000). For the annotation of the coreference relations in the KNACK-2002 corpus, we used MITRE's "Alembic Workbench" as annotation environment[3]. The following is an example of such an annotated piece of text:

(2) Ongeveer een maand geleden stuurde <COREF ID = "1"> American Airlines </COREF> <COREF ID = "2" MIN = "toplui"> enkele toplui </COREF> naar Brussel. <COREF ID = "3" TYPE = "IDENT" REF = "1" MIN="vliegtuigmaatschappij"> De grote vliegtuigmaatschappij </COREF> had interesse voor DAT en wou daarover <COREF ID = "5"> de eerste minister </COREF> spreken. Maar <COREF ID = "6" TYPE = "IDENT" REF = "5"> Guy Verhofstadt </COREF> (VLD) weigerde <COREF ID = "7" TYPE = "BOUND" REF = "2"> de delegatie </COREF> te ontvangen.

English: About one month ago, American Airlines sent some senior executives to Brussels. The large airplane company was interested in DAT and wanted to discuss the matter with the prime minister. But Guy Verhofstadt (VLD) refused to see the delegation.

In (2), three coreference chains (sequences of NPs referring to each other) are marked: one for *"American Airlines"* and *"De grote vliegtuigmaatschappij"*, a second chain with *"enkele toplui"* and *"de delegatie"* and a third chain with *"de eerste minister"* and *"Guy Verhofstadt"*. The annotation of this example sentence and all other sentences in our Dutch corpus mainly follows the MUC-7 guidelines (MUC-7 1998). As in the MUC annotations, all coreferences start with a <COREF> tag and are closed with a </COREF> close tag. The initial <COREF> tag contains additional information about the coreference: the unique ID of the NP (ID), the type of coreference relation (TYPE), the ID of the entity

[2] The annotation scheme is available at http://www.cnts.ua.ac.be/~hoste/proefschrift/AppendixA.pdf.

[3] More information on this workbench can be found at http://www.mitre.org/tech/alembic-workbench.

referred to (REF) and optionally the minimal tag of the coreference (MIN). For a detailed description of the annotated relations, we refer to Hoste (2005).

In total, the KNACK-2002 corpus consists of 267 documents annotated with coreference information. In this corpus, 12,546 noun phrases are annotated with coreferential information. Not only did this annotation effort enable us to assess the difficulty of the task, it also led to a corpus which can be used for the evaluation and the development of different approaches to automatic coreference resolution for Dutch.

3 Data preparation

For the experiments, we made a random, but balanced selection of 50 documents covering different topics. We selected 10 documents covering internal politics, 10 documents on foreign affairs, another 10 documents on economy, 5 documents on health and health care, 5 texts covering scientific topics and finally 10 documents covering a variety of topics (such as sports, education, history and ecology). In total, the documents contain 25,994 words and 3,014 coreferential tags. Half of the texts was used as training set and the other half as test set. The division between testing and training material was done randomly at document level (in order to avoid documents being divided in two). The KNACK-2002 training and test set contain 1,688 and 1,326 coreferential NPs, respectively.

3.1 Preprocessing

For the construction of the data sets, we selected all noun phrases in the KNACK-2002 corpus. These noun phrases could be detected after preprocessing the raw text corpora. The following preprocessing steps were taken: tokenization by means of a rule-based system using regular expressions, named entity recognition using the memory-based learning approach of De Meulder and Daelemans (2003), part-of-speech tagging, text chunking and relation finding (all three modules trained on the Spoken Dutch Corpus (CGN)[4] as described in Tjong Kim Sang, Daelemans and Höthker (2004)). We also performed a machine learned morphological analysis (De Pauw, Laureys, Daelemans and Van hamme 2004).

3.2 Instance construction

On the basis of the preprocessed texts, we selected positive and negative instances for the training data and test instances for the test data.

Positive and negative instances As exemplified in Table 1, the positive instances were made by combining each anaphor with each preceding element in the coreference chain. The negative instances were built (i) by combining each anaphor with each preceding NP which was not part of any coreference chain and (ii) by combining each anaphor with each preceding NP which was part of another

[4] More information on this corpus can be found at http://lands.let.ru.nl/cgn/

coreference chain. In order to reduce the number of negative training instances, we restricted the search scope to 20 sentences preceding the candidate anaphor. This instance construction led to a training instance base of 102,376 instances for the 1,687 references in the training data.

Table 1: Training instance construction for the pronoun “ze” as given in example (1).

ze	een bescheiden bedrag	neg
ze	43 miljoen gulden losgeld	neg
ze	de kidnappers	pos
ze	zijn chauffeur	neg
ze	zijn	neg
ze	Alfred Heineken	neg
ze	9 november 1983	neg

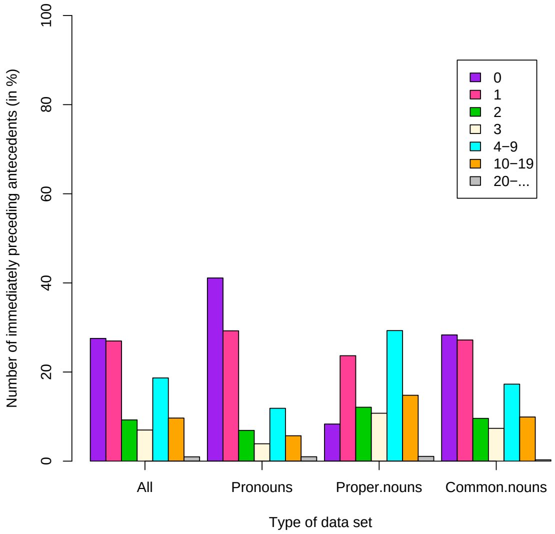
Test instances For the construction of the test instances, all NPs starting from the second NP in a text are considered a possible anaphor, whereas all preceding NPs are considered possible antecedents. Since this type of instance construction leads to an enormous increase of the data set and since we are eventually only interested in finding one possible antecedent per anaphor, we took into account some search scope limitations.

As a starting point for restricting the number of instances without losing possibly interesting information, we calculated the distance between the references and their immediately preceding antecedent in the training data. The distances were calculated as follows: antecedents from the same sentence as the anaphor were at distance 0. Antecedents in the sentence preceding the sentence of the referring expression, were at distance 1, and so on. We divided the group of referring expressions into three categories: (1) pronouns, (2) proper nouns and (3) common nouns. These results are displayed in Figure 1. It shows that for the pronouns 77.3% of the immediately preceding antecedents can be found in a context of three sentences. With respect to the named entities, we can observe that 44.0% of the immediately preceding antecedents can be found in a scope of three sentences. For the common noun NPs, this percentage is 65.2%. We used this information in the construction of the test instances. For the pronouns, all NPs in a context of 2 sentences before the pronominal NP were included in the test sets for the pronouns (as for example also in Yang, Zhou, Su and Tan (2003) for English). For the proper and common nouns, all partially matching NPs were included. For the non matching NPs, the search scope was restricted to two sentences. This instance selection allowed us to obtain an overall test set reduction.

3.3 One vs. three

Instead of merging the different types of NPs into one single training and test set (as for example Ng and Cardie (2002) and Soon, Ng and Lim (2001) for English),

Figure 1: Distance in number of sentences between a given referring expression and its immediately preceding antecedent in the KNACK-2002 training set.



we built 3 smaller datasets. This resulted in a learning system for pronouns, one for named entities and a third system for the other NPs. The main motivation for this approach is that other information sources play a role in the resolution of pronominal references than for example in the resolution of references involving proper nouns. Example sentence (3) clearly shows the importance of string matching or aliasing in the resolution of proper nouns. These features are less important for the resolution of the coreferential link between a pronoun and a common noun NP in example (4), for which information on gender, number and distance is crucial.

(3) **Vlaams minister van Mobiliteit Steve Stevaert** dreigt met een regeringscrisis als de federale regering blijft weigeren mee te werken aan het verbeteren van de verkeersveiligheid. (...) **Stevaert** ergert zich aan de manier waarop de verschillende ministeries het dossier naar elkaar toeschuiven.

(4) **De beklaagde**, die de doodstraf riskeert, wil dat **zijn** proces op televisie uitgezonden wordt.

The resulting data sets are displayed in Table 2. The ‘Pronouns’ data set contains the NPs ending on a personal, reflexive or possessive pronoun. The ‘Proper nouns’ data set contains the NPs which have a proper noun as head, whereas the ‘Common nouns’ data set contains all other NPs which are not in the two other categories. And the fourth dataset is the sum of all three datasets. This grouping of the different types of NPs does not only allow for building more specialized classifiers, it also makes error analysis more transparent (as shown in section 7).

Table 2: Number of instances per NP type in the KNACK-2002 corpus.

	TRAIN		TEST
NP type	positive	negative	
Pronouns	3,111	33,155	5,897
Proper nouns	2,065	31,370	10,954
Common nouns	1,281	31,394	24,677
Complete	6,457	95,919	41,528

4 Selection of informative features

Several information sources contribute to a correct resolution of coreferential relations, viz. morphological, lexical, syntactic, semantic and positional information and also world-knowledge. In this section, we give an overview of the information sources we used for the construction of the instances. These are so-called shallow information sources, namely information sources which are easy to compute.

- The **positional features** give information on the location of the candidate anaphors and antecedents. We use the following three positional features: DIST_SENT (giving information on the number of sentences between the candidate anaphor and its candidate antecedent), DIST_NP (giving information on the number of noun phrases between the candidate anaphor and its candidate antecedent) and the binary feature DIST_LT_THREE (which is set to 'yes' if both constituents are less than three sentences apart from one another and 'no' if both constituents are more than three sentences apart).

- The **local context features** inform on the three words preceding and following the candidate anaphor, with their corresponding part-of-speech tags.

- As **morphological and lexical features**, the I_PRON, J_PRON and I+J_PRON features indicate whether a given candidate anaphor, its candidate antecedent or both are pronouns (personal, possessive, demonstrative or reflexive). The feature J_PRON_I_PROPER indicates whether the possible antecedent of a coreferential pronoun is a proper noun. J_DEMON and J_DEF give information on the demonstrativeness and definiteness of the candidate anaphor. I_PROPER, J_PROPER and BOTH_PROPER indicate whether a given candidate anaphor, its candidate antecedent and both are proper names. And finally, NUM_AGREE looks for number agreement between the candidate anaphor and its candidate antecedent.

- The **syntactic features** ANA_SYNT and ANT_SYNT inform on the syntactic function (subject, object, predicate) of the candidate anaphor and its antecedent. If the candidate antecedent is the immediately preceding subject, object or predicate, it takes as value 'imm_prec_SBJ', 'imm_prec_OBJ' or 'imm_prec_PREDC', respectively. The BOTH_SBJ/OBJ feature checks for syntactic parallelism. The APPOSITIVE feature checks whether the coreferential NP is an apposition to the preceding NP.

- As **string-matching features**, the following features were used: COMP_MATCH, which checks for a complete match between the anaphor and its candidate antecedent and the PART_MATCH feature, which checks for a partial match between both noun phrases. We also performed word internal matching. In order to do so, we used the previously described morphological analysis to split the compound words into their different parts, e.g. "pensioenspaarverzekeringen" into "pensioen+spaar+verzekeringen" and "pensioenverzekeringen" into "pensioen+verzekeringen'. These different parts were then checked for partial matching. Furthermore, the ALIAS feature indicates whether the candidate anaphor is an alias of its candidate antecedent or vice versa. The alias of a given NP is determined by removing all prepositions and determiners and then by taking the first letter of the nouns in the noun phrase. These letters are then combined in various ways. This simple approach allows us to capture the alias "IBM" which stands for "**I**nternational **B**usiness **M**achines". Finally, the SAME_HEAD feature checks whether the anaphor and its candidate antecedent share the same head. An example of two NPs sharing the same head is "de Golf" and "de Perzische Golf".

- For the extraction of the **semantic features** for the proper nouns , we took into account lists with location names, organization names, person names and male and female person names. Lacking this type of information for the common noun NPs, we used the Celex lexical data base (Baayen, Piepenbrock and van Rijn 1993) instead to provide gender information for the head nouns of the common noun NPs. There are three basic genders in Dutch: male, female and neutral. In addition, CELEX also names female nouns which can be treated as male and nouns whose gender depends on the context in which they are used. This makes five feature values with gender information: 'male', 'female', 'neutral', 'female(male)', 'male-female'. For the extraction of the SYNONYM and HYPERNYM feature, we used all synonyms and hypernyms in the Dutch EuroWordNet (http://www.illc.uva.nl/EuroWordNet) output. And finally, SAME_NE makes use of the output of the Dutch named entity recognition system described earlier.

5 A machine learning approach

5.1 A lazy and an eager learner

Having built the feature vectors for our experiments, we can now continue with a description of the machine learning approaches which we used for our experiments. For the experiments, two machine learning packages were used: the memory-based learning package TIMBL (Daelemans, Zavrel, van der Sloot and van den Bosch 2002)[5], and the rule induction package RIPPER (Cohen 1995). Both TIMBL and RIPPER require as input an example represented as a vector of real-valued or symbolic features, followed by a class. But the two learning methods have a completely different 'bias'. They use different search heuristics and behave differently in the way they represent the learned knowledge.

The first learning approach applied to our coreferentially annotated data, is a *memory-based learning (MBL)* approach. For our experiments, we used the memory-based learning algorithms implemented in TIMBL (Daelemans et al. 2002). During learning MBL keeps all training data in memory and at classification time, a previously unseen test example is presented to the system and its similarity to all examples in memory is computed using a similarity metric. The class of the most similar example(s) is then used as prediction for the test instance. This strategy is often referred to as "lazy" learning. This storage of all training instances in memory during learning, without abstracting and without eliminating noise or exceptions is the distinguishing feature of memory-based learning in contrast with minimal-description-length-driven or "eager" ML algorithms (e.g. decision trees, rules and decision lists).

The second learning method used in our experiments is the rule learning system RIPPER, which has been developed by Cohen (1995). During learning, RIPPER induces classification rules on the basis of the set of preclassified examples. This type of learning approach is called an eager learning approach, since there is a compression of the training material into a limited number of rules.

5.2 A two-step procedure

The general setup of our experiments is the following. Both RIPPER and TIMBL are trained on the complete training set and the resulting classifiers are applied to the held-out test set, which is represented as a set of instances. Defining the coreference resolution process as a classification problem, however, involves the use of a two-step procedure. In a **first step**, the classifiers are cross-validated on the training data. For this first step, we performed an extensive optimization through feature selection, the optimization of the algorithm parameters and through different sampling techniques in order to have a more balanced class distribution (see Hoste (2005) for a detailed description of this optimization procedure). These optimized classifiers then decide on the basis of the information learned from the training set whether the combination of a given candidate coreference and its can-

[5] Available from http://ilk.uvt.nl

didate antecedent in the test set is classified as a coreferential link. Since each NP in the test set is linked with several preceding NPs, this implies that one single coreference can be linked to more than one antecedent, which for its part can also refer to multiple antecedents, and so on. Therefore, a **second step** is taken, which involves the selection of one coreferential link per coreference. In this second step, the coreferential chains are built on the basis of the positively classified instances. We can illustrate this procedure for the coreferential relation between "hij" and "President Bush" in example (5).

(5) **President Bush** heeft Verhofstadt ontmoet in Brussel. **Hij** heeft met onze eerste minister de situatie in Irak besproken.
English: **President Bush** met Verhofstadt in Brussels. **He** spoke with our prime minister about the situation in Iraq.

Table 3: Test instances built for the "hij" in example (5)

Antecedent	Coreference	Classification
Brussel	hij	no
Verhofstadt	hij	yes
President Bush	hij	yes

For the NP "hij" test instances are built for the NP pairs displayed in Table 3. The result of the first step might be that the learner classifies the first instance as non-coreferential and the last two instances as being coreferential. Since we only want to select one antecedent per anaphor, a second step is taken to make a choice between the two positive instances (hij - Verhofstadt) and (hij - President Bush). For this **second step**, different directions can be taken. The most straightforward approach is to take as antecedent the first NP found to be coreferent with the anaphor (as in Soon et al. (2001)). Other approaches (Ng and Cardie 2002, Yang et al. 2003) assign scores to the candidate antecedents and select the most likely antecedent among the candidate antecedents. This is also the antecedent selection strategy we have taken (see Hoste (2005) for more information).

5.3 Evaluation procedure

We will report performance in terms of precision, recall and F-measure, using the MUC scoring program from Vilain, Burger, Aberdeen, Connolly and Hirschman (1995). The program looks for the evaluation at equivalence classes, being the transitive closure of a coreference chain. In the Vilain et al. (1995) algorithm, the **recall** for an entire set T of equivalence classes is computed as follows:

$$R_T = \frac{\sum(c(S)-m(S))}{\sum(c(S))}$$

where $c(S)$ is the minimal number of correct links necessary to generate the equivalence class S: $c(S) = (|S| - 1)$. $m(S)$ is the number of missing links in the response relative to equivalence set S generated by the key: $m(S) = (|p(S)| - 1)$.

$p(S)$ is a partition of S relative to the response: each subset of S in the partition is formed by intersecting S and the responses sets R_i that overlap S. For the computation of the **precision**, the roles for the answer key and the response are reversed. For example, equivalence class S can consist of the following elements S = $\{1\ 2\ 3\ 4\}$. If the response is $< 1 - 2 >$, then $p(S)$ is $\{1\ 2\}$, $\{3\}$ and $\{4\}$.

6 Experimental results

In order to evaluate the performance of our classifiers, we first calculated two baseline scores.

6.1 Two baseline scores

- **Baseline I**: For the calculation of the first baseline, we did not take into account any linguistic, semantic or location information. This implies that this baseline is calculated on the large test corpus which links every NP to every preceding NP and not on the smaller test corpora described in Section 3 which already take into account feature information. Baseline I is obtained by linking every noun phrase to its immediately preceding noun phrase.

- **Baseline II**: This somewhat more sophisticated baseline is the result of the application of some simple rules: select the closest antecedent with the same gender and number (pronouns), select the closest antecedent which partially/completely matches the NP (proper and common nouns).

Table 4: Two baseline scores. The recall and $F_{\beta=1}$ scores could not be provided for the NP type data sets, since the scoring software does not distinguish between the three NP types.

		Prec.	Rec.	$F_{\beta=1}$
Baseline I	PPC	27.9	**81.9**	41.7
	Pronouns	18.1	—	—
	Proper nouns	2.4	—	—
	Common nouns	4.9	—	—
Baseline II	PPC	**38.9**	45.7	**42.0**
	Pronouns	39.2	—	—
	Proper nouns	56.9	—	—
	Common nouns	23.6	—	—

Table 4 shows the precision, recall and $F_{\beta=1}$ scores for these two baselines. The errors associated with these measures can be interpreted as follows. The recall errors are caused by classifying positive instances as being negative. These false negatives cause missing links in the coreferential chains. The precision errors, on the other hand, are caused by classifying negative instances as being positive. These false positives cause spurious links in the coreferential chains. Table 4 reveals the following tendencies. Linking every NP to the immediately preceding NP, as was done for the first baseline, leads to a high overall recall score of 81.9%, whereas the precision is low: 27.9%. The Baseline II scores which depend on feature information, are more balanced: 45.7% recall and 38.9% precision. The highest $F_{\beta=1}$ value is obtained by Baseline II: 42.0%. With respect to the baseline results on the NP type data sets, the following observations can be made. The Baseline I results are low, except for the precision scores for the pronouns (18.1%). This result confirms that the antecedent of a pronominal anaphor is located close to the anaphor, as already shown in Section 3.

6.2 Classifier results

Table 5 gives an overview of the results obtained by TIMBL and RIPPER in terms of precision, recall and $F_{\beta=1}$. Table 5 shows that both TIMBL and RIPPER obtain an overall $F_{\beta=1}$ score of 51.0%. The precision scores for the "Pronouns" (64.9% for TIMBL and 66.7% for RIPPER) and the "Proper nouns" data sets (79.4% for TIMBL and 79.0% for RIPPER) are much higher than those obtained on the "Common nouns" data set (47.6% for TIMBL and 47.5% for RIPPER). Furthermore, the recall scores are about 20% lower than the precision scores, which implies that most of the errors represent missing links: 42.2% recall vs. 65.9% precision for TIMBL and 40.9% recall vs. 66.3% precision for RIPPER. Overall, we can conclude from these results that coreference resolution for Dutch still presents some major challenges.

As a test of the methodology used all experiments were also performed on the widely used English MUC-6 and MUC-7 data sets, for which state-of-the art results could be reported: 64.3% (TIMBL) and 63.4% (RIPPER) for MUC-6 and 60.2% (TIMBL) and 57.6% (RIPPER) for MUC-7. For an elaborate description of experiments on these data sets, we refer to Soon et al. (2001), Ng and Cardie (2002) and Hoste (2005).

Table 5: Results from TIMBL and RIPPER in terms of precision, recall and $F_{\beta=1}$. No recall and $F_{\beta=1}$ scores could be provided on the NP type data sets, since the scoring software does not distinguish between the three NP types.

		Prec.	Rec.	$F_{\beta=1}$
Timbl	PPC	65.9	42.2	51.4
	Pronouns	64.9	—	—
	Proper nouns	79.4	—	—
	Common nouns	47.6	—	—
Ripper	PPC	66.3	40.9	50.6
	Pronouns	66.7	—	—
	Proper nouns	79.0	—	—
	Common nouns	47.5	—	—

7 Error analysis

Although we cannot quantify the different types of errors, since this would require a manual analysis of the complete test corpus, we performed a qualitative error analysis on three KNACK-2002 documents. We selected one document on which our system performs above average and two documents for which the $F_{\beta=1}$ score is below average. In each of these documents, we looked for the errors committed by the different learning modules. We will now discuss these errors and some directions for future research.

Pronouns The main source of errors for the pronominal resolution system is the lack of features which can capture the pleonastic and coreferential use of pronouns (as in 7). Therefore, more effort should be put in features which can capture this difference. Another possible approach is to train a classifier, as in Mitkov, Evans and Orasan (2002), which automatically classifies instances of "it" as pleonastic or anaphoric. The resolution of the pronominal anaphors is also hindered by part-of-speech tagging errors (as in 6). E.g. the female "ze" is often erroneously tagged as a third person plural pronoun and vice versa. Furthermore, for the Dutch male and female pronouns, such as "hij", "hem", "haar", the search space of candidate antecedents is much larger than that for the corresponding English pronouns, since they can also refer to the linguistic gender of their antecedent, as shown in (8).

(6) **De moeder van Moussaoui** gaf een persconferentie waarin **ze** om een eerlijk proces vroeg.
English: **The mother of Moussaoui** gave a press conference in which **she** asked for a fair trial. (Missing link)

(7) Een god van **het vuur**. Paul Wolfowitz heeft alles bij elkaar eigenlijk een bescheiden job in de Amerikaanse regering. Hoe komt **het** dan dat hij zoveel invloed heeft in het Witte Huis?
English: A god of **the fire**. In the end, Paul Wolfowitz has a rather insignificant job in the American government. How is **it** possible that he has so much influence in the White House? (Spurious link)

(8) Zij stelden dat het moeilijk zou zijn om **de studie** te 'dupliceren'. Waarmee werd gezegd dat **ze** niet wetenschappelijk was uitgevoerd.
English: They argued that it would be hard to 'duplicate' **the study**. By which was claimed that **it (Dutch: "she")** was not carried out in a scientific way. (Missing link)

Proper nouns Although high recall and precision scores can be observed for the proper nouns, there is still room for improvement. The errors are mainly caused by preprocessing errors: errors in NP chunking and errors in part of speech tagging (9), etc. The part-of-speech tagger trained on the Spoken Dutch Corpus mainly assigns three different types of tags to proper nouns: SPEC(deeleigen) (as for "Zacarias Moussaoui", SPEC(afgebr) (as for "Moussaoui") and "N(eigen (...)". The corresponding chunks for the underlying part-of-speech tags are "MWU" (multi word unit) for SPEC(deeleigen) and SPEC(afgebr) and "NP" for "N(eigen (...)". Since multi word units can also consist of non-NP combinations (e.g. "in staat"), these multi word units are not always selected for resolution.

(9) **Zacarias Moussaoui**, de eerste persoon die door het Amerikaanse gerecht aangeklaagd is (...) De moeder van **Moussaoui** vloog enige dagen voor zijn voorleiding naar de Verenigde Staten.
English: **Zacarias Moussaoui**, the first person who has been charged by the American judicial authorities (...) The mother of **Moussaoui** came to the United States a few days before the hearing. (Missing link)

Common nouns As also observed for other languages (e.g. Ng and Cardie (2002), Hoste (2005) for English and Strube, Rapp and Müller (2002) for German), the resolution of coreferential relations between common noun NPs is problematic. As for the resolution of coreferential proper nouns and pronouns, missing links can be caused by preprocessing errors, such as errors in part-of-speech tagging, NP chunking, apposition recognition, etc. We therefore conclude that the shallow parser trained on the Spoken Dutch Corpus might not be suitable for this text corpus. We therefore plan to reconsider the whole preprocessing procedure. Other errors are typical for the resolution of coreferential relations between common nouns: the lack of recognizing synonyms as in (10), the lack of recognizing hyponyms as in (11), and the lack of world knowledge (11). For the construction of the semantic features, we used named entity recognition and the Dutch EuroWordNet. But this lexical resource is very restricted and misses a lot of commonly used expressions and their lexical relations. Furthermore, a lot of coreferential relations are restricted in time, such as the pair "Chirac"-"the president of France", or names of political parties (e.g. "de groenen"-"Agalev"-"Groen!"). In order to overcome this lack of information in the existing resources and in order to capture "dynamic" coreferential relations, we plan to use the Web as a resource (as for example Keller, Lapata and Ourioupina (2002) and Modjeska, Markert and Nissim (2003)).

(10) Stevaert en Charles Picqué gaven elkaar de schuld voor het disfunctioneren van **twee onbemande camera's** op de A12. Picqué - bevoegd voor de erkenning van **de flitspalen** - (...)
English: Stevaert and Charles Picqué blamed each other for the disfunctioning of **two unmanned cameras** at the A12. **Picqué** - authorized for the homologation of **the flash-guns** - (...) (Missing link)

(11) **Zacarias Moussaoui**, de eerste persoon die aangeklaagd is voor **de terreuraanvallen van 11 september**, pleit onschuldig bij zijn eerste verschijning voor de rechtbank. (...) **De Fransman van Marokkaanse afkomst** wordt ervan verdacht de 'twintigste vliegtuigkaper' te zijn die door omstandigheden niet aan **de kapingen** kon deelnemen.
English: Zacarias Moussaoui, the first person who has been charged for **the terrorist attacks of 11 September**, pleads not guilty at the first hearing. (...) **The French citizen of Moroccan descent** is accused of being the 'twentieth hijacker' who was prevented from carrying out **the hijackings**. (Missing link)

8 Summary

In this paper, we presented a machine learning approach to the resolution of coreferential relations between nominal constituents in Dutch. It is the first corpus-based resolution approach proposed for this language. The corpus-based strategy was enabled by the annotation of a corpus with coreferential information for pronominal, proper noun and common noun coreferences: KNACK-2002.

The F scores of 51% of both TIMBL and RIPPER on the held-out test data and the qualitative error analysis showed that coreference resolution for Dutch presents some major challenges. Especially the resolution of coreferential links between common noun NPs is problematic and suffers from lacking semantic and world knowledge. Similar observations could be made for the English MUC-6 and MUC-7 data sets.

References

Baayen, R., Piepenbrock, R. and van Rijn, H.(1993), The celex lexical data base on cd-rom.

Baldwin, B.(1997), Cogniac: high precision coreference with limited knowledge and linguistic resources, *Proceedings of the ACL'97/EACL'97 workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pp. 38–45.

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

Cohen, W. W.(1995), Fast effective rule induction, *Proceedings of the 12th International Conference on Machine Learning (ICML-1995)*, pp. 115–123.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A.(2002), Timbl: Tilburg memory-based learner, version 4.3, reference guide, *Technical Report ILK Technical Report - ILK 02-10*, Tilburg University.

Davies, S., Poesio, M., Bruneseaux, F. and Romary, L.(1998), Annotating coreference in dialogues: Proposal for a scheme for mate, http://www.hcrc.ed.ac.uk/˜poesio/MATE/anno_manual.htm.

De Meulder, F. and Daelemans, W.(2003), Memory-based named entity recognition using unannotated data, *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pp. 208–211.

De Pauw, G., Laureys, T., Daelemans, W. and Van hamme, H.(2004), A comparison of two different approaches to morphological analysis of dutch, *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pp. 62–69.

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

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

Hoste, V.(2005), *Optimization Issues in Machine Learning of Coreference Resolution*, PhD thesis, Antwerp University.

Keller, F., Lapata, M. and Ourioupina, O.(2002), Using the web to overcome data sparseness, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pp. 230–237.

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

Mitkov, R., Evans, R. and Orasan, C.(2002), A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method, *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*.

Modjeska, N., Markert, K. and Nissim, M.(2003), Using the web in machine learning for other-anaphora resolution, *Proceedings of the 2003 Conference on Empirical Methods in Natural Lanugage Processing (EMNLP-2003)*, pp. 176–183.

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

Ng, V. and Cardie, C.(2002), 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)*, pp. 55–62.

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

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

Strube, M., Rapp, S. and Müller, C.(2002), The influence of minimum edit distance on reference resolution, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pp. 312–319.

Tjong Kim Sang, E., Daelemans, W. and Höthker, A.(2004), Reduction of dutch sentences for automatic subtitling, *Computational Linguistics in the Netherlands 2003. Selected Papers from the Fourteenth CLIN Meeting*, pp. 109–123.

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

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

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