Disambiguation of the Neuter Pronoun and its Effect on Pronominal Coreference Resolution

Véronique Hoste[1], Iris Hendrickx[2], and Walter Daelemans[2]

[1] LT3 - Language and Translation Technology Team,
University College Ghent, Groot-Brittaniëlaan 45, Ghent,
Belgium
veronique.hoste@hogent.be
[2] CNTS - Language Technology Group,
University of Antwerp, Universiteitsplein 1, Antwerp,
Belgium
iris.hendrickx@ua.ac.be, walter.daelemans@ua.ac.be

Abstract. Coreference resolution, determining the appropriate discourse referent for an anaphoric expression, is an essential but difficult task in natural language processing. It has been observed that an important source of errors in machine-learning based approaches to this task, is the wrong disambiguation of the third person singular neuter pronoun as either referential or non-referential. In this paper, we investigate whether a machine learning based approach can be successfully applied to the disambiguation of the neuter pronoun in Dutch and show a modest potential effect of this disambiguation on the results of a machine learning based coreference resolution system for Dutch.

1 Introduction

Coreference resolution, the task of determining the appropriate discourse referent for a given anaphoric expression, has gained increasing popularity in natural language processing research and has become a key component in applications such as information extraction, question answering, automatic summarization, etc. in which text understanding is of major importance.

In this paper we focus on pronominal coreference resolution, and more specifically on the improvement of a machine learning system for automatic pronominal coreference resolution through the automatic disambiguation of "het" (Eng.: "it") as either referential or non-referential. The focus is on the classification of the Dutch neuter "het", in contrast to most of the related work which is mainly oriented towards English. In order to classify the third person singular neuter pronoun, two different types of approaches have been proposed for English: rule-based strategies ([1], [2]) and machine learning approaches ([3], [4]).

Although the existing approaches to the automatic identification of the different uses of the third person singular neuter pronouns are always motivated by the task of pronominal coreference resolution, this effect of the automatic classification of "it" on resolution performance has to our knowledge not yet been

investigated, except by [5] who performed a more global comparison of resolution perfomance with and without the detection of non-anaphoric constituents. The goal of this paper is twofold. In a first step, we investigate whether the distribution of the different uses of the Dutch neuter pronoun is similar to the one reported for English. Furthermore, we investigate whether the machine learning approach, successfully applied for the English "it", can be easily ported to the automatic classification of the Dutch "het". In a second step, we evaluate the effect of this classification on a learning approach for Dutch pronominal coreference resolution as described in [6]. Since the coreference resolution system is designed to detect coreferential chains between nominal constituents, we are mainly interested in the detection of the pronouns referring to antecedent NPs.

The remainder of this paper is organized as follows. Section 2 introduces the data used for the experiments and describes the annotation and the distribution of phenomena of interest compared to English data used for the same task. The experimental set-up and results are described in Section 3. The effect of the separate disambiguation component on overall anaphora resolution is discussed in Section 4, and Section 5 summarizes the main findings of the paper.

2   Data Sets

For the experiments, the focus was on Dutch coreference resolution. Two corpora were annotated with information on the third person neuter pronoun: KNACK, a corpus of news magazine texts (106,011 tokens) and SPECTRUM, a corpus with medical encyclopedia texts (133,887 tokens). Two linguists annotated the corpora in parallel in accordance with the annotation guidelines described below, which are based on the general Dutch grammar (ANS)[3]. As input, the annotators received free text in which all occurrences of "het" were marked, the majority of which involved "het" as definite article. For the annotation of the personal pronoun "het", the annotators had to differentiate between the non-referential use of the pronoun as in example (1) and its referential use. In example (1), "het" is part of an idiomatic expression and not referential.

   (1)   Leopold haalt scherp uit naar onder meer Hubert Pierlot, de eeuwige zondebok
         met wie hij **het** niet kon vinden.
         English: Leopold sharply attacks among others Pierlot, the eternal scapegoat
         with whom he can't get on.

We distinguished between the following four types of referential use: (i) reference to preceding "het" words as in example (2), (ii) reference to a preceding clause as in example (3), (iii) "het" as anticipatory subject (4) and finally, "het" as subject of a nominal predicate (5).

   (2)   Weet je waar **mijn boek** is? Ik heb **het** niet gezien.
         English: Do you know where **my book** is? I haven't seen **it**.

---

(3) **Leopold III kwam aan de macht nadat zijn vader in 1934 was veron-
gelukt in Marche-les-Dames.** Volgens historicus Jan Verwelkenhuyzen ging
toen het gerucht dat de Duitsers **het** zo hadden gewild.
English: **Leopold III came to power after his father died in an acci-
dent in 1934 in Marche-les-Dames.** According to historian Jan Verwelken-
huyzen, there was a rumour that the Germans had wanted **it** that way.

(4) **Het** lijkt er namelijk op **dat de bevolking van Zimbabwe haar huis-
bakken dictator meer dan beu is**.
English: **It** seems **the population of Zimbabwe has had it with its home-
grown dictator**.

(5) **Het** zijn, voorlopig althans, **slechts schuchtere signalen**.
English: **It** is, for now, **only a weak signal**.

On the Knack data, a kappa agreement score was obtained of 0.74; on the
Spectrum data, the kappa score was 0.81. After this first annotation round,
both annotators re-annotated the texts jointly in order to reach a consensus
annotation. In total, 6560 occurrences of "het" were annotated, of which 844 are
pronominal. Table 1 gives an overview of the distribution of the different uses
of the neuter pronoun in both annotated Dutch corpora. For English, we also
provided the number of times the "it" refers to a preceding noun phrase. The
table reveals that the English corpora which have been previously used for the
automatic classification of "it" all show a large number (>67%) of occurrences
of "it" in which the pronoun refers to a preceding noun phrase. For Dutch,
however, this percentage drops to around 20% for the newspaper texts, whereas
for the medical texts nearly half of the "het" occurrences refer to an NP. Taking
the Dutch corpora as a whole, three categories show a similar distribution: the
non-referential use (30.1%), the reference to a preceding noun phrase (32.6%)
and the neuter pronoun as anticipatory subject (23.3%).

Table 1. Distribution of the pronominal "het" in the different data sets.

| | DUTCH | | | ENGLISH |
	Knack	Spectrum	Total	
Pronominal use	507/2890	337/3670	884/6560	
Non-referential	39.0%	17.7%	30.1%	
Ref - preceding clause	5.7%	0.3%	3.5%	
Ref - noun phrase	**21.3%**	**49.5%**	**32.6%**	MUC-6 **74.4%**
				MUC-7 **80.7%**
				[4] **67.9%**
				[3] **69.6%**
Ref - anticipatory subject	19.9%	28.5%	23.3%	
Ref - anticipatory object	5.1%	1.2%	3.5%	
Ref - nominal predicate	8.9%	3.9%	6.9%	

3 Experimental Setup

For the construction of the machine learning data sets, the following preprocessing steps were taken. Lemmatization was performed using a memory-based lemmatizer trained on a lexicon derived from the Spoken Dutch Corpus (CGN)[4], a 10-million word corpus of spoken Dutch. Part-of-speech tagging and text chunking were performed by the memory-based tagger MBT[7], which was also trained on the CGN corpus. For all occurrences of "het", a feature vector was built consisting of 38 features. These features include positional information (sentence number and position in sentence), information on the focus word itself (wordform, part-of-speech and chunk information), furthermore information on the wordform, lemma and part-of-speech of five words before and after the focus word, and finally information on the use of a preposition before the focus word [1]. Based on the assumption that verbs which occur more often with "het" indicate the non-anaphoric use of the pronoun, we included a last feature for which the association strength was calculated between "het" as a subject and its accompanying verb. This association strength was represented by mutual information scores and was based on the Dutch Twente News Corpus (500 million words). A minimal cut-off frequency of 1000 was chosen.

For the classification of the different uses of "het", we used a memory-based learning algorithm, as was also previously applied to this task by [4] and [3]. Memory-based learning (a k-nearest neighbor approach) is a lazy learning approach that stores all training data in memory. At classification time, the algorithm classifies new instances by searching for the nearest neighbors to the new instance using a similarity metric, and extrapolating from their class. In our experiments we use the TIMBL [7] software package[5] that implements a version of the k-nn algorithm optimised for working with linguistic datasets and that provides several similarity metrics and variations of the basic algorithm. Since these different parameters, individually and in combination, can strongly affect the functioning of the algorithm, we performed joint feature selection and parameter optimization by means of a generational genetic algorithm as described in [6] and as shown in Figure 1. Given the modest size of the data sets, leave-one-out was used for validation. The following parameters were varied: the number of nearest neighbours, expressed by k, the distance metric and the model to extrapolate from the nearest neighbours. For the three data sets, viz. Knack, Spectrum and the concatenation of the two, optimization led to a selection of a high k value (9 for Spectrum and the concatenated data; 16 for Knack) and to the selection of exponential decay distance weighted voting and of gain ratio (a normalised version of information gain) as distance metric for the three data sets. Feature selection led to an omission of the feature informing on the position of the word in the sentence and to a selection in the local context features. For the Knack data, the association strength feature was also filtered out.

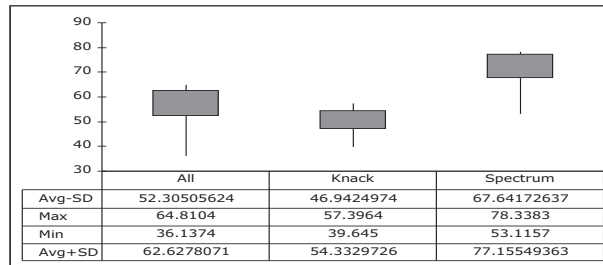|	All	Knack	Spectrum
Avg-SD	52.30505624	46.9424974	67.64172637
Max	64.8104	57.3964	78.3383
Min	36.1374	39.645	53.1157
Avg+SD	62.6278071	54.3329726	77.15549363

Fig. 1. Optimization results for the three data sets. The graphs show the difference between the best and the worst parameter and feature subset combination per data set. The boxes in the graphs represent averages and deviations.

Tables 2 gives an overview of the overall and 5-ary classification results of the optimized memory-based classifier. It shows a 30% improvement over the most frequent sense accuracy for the three data sets. The results also show that for some subtypes of referential use, there is too little evidence in the training data to train an accurate classifier on. The non-referential use of "het", on the other hand, can be detected with a reliability of >70%. For the "het" which refers to a preceding noun phrase, divergent F-scores are obtained: 39.1% for Knack, as opposed to 83.4% for Spectrum.

Table 2. Performance in terms of accuracy, precision, recall and F-score of TIMBL on the three data sets. Both the overall and 5-ary classification results of the optimized memory-based classifier are given. As baseline score, the most frequent class, i.e. reference to a preceding NP, was taken and kept constant over all data sets.

	Knack			Spectrum			Total		
Baseline	21.3			49.5			32.6		
Accuracy	57.40			78.34			64.81		
	P	R	F	P	R	F	P	R	F
Non-referential	60.20	89.39	71.95	85.11	71.43	77.67	67.49	75.20	71.14
Ref - preceding clause	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ref - noun phrase	49.30	32.41	39.11	78.72	88.62	83.38	63.16	69.82	66.32
Ref - anticipatory subject	52.34	55.45	53.85	74.51	79.17	76.77	62.77	73.60	67.76
Ref - anticipatory object	100.00	42.31	59.46	0.00	0.00	0.00	100.00	23.33	37.84
Ref - nominal predicate	54.55	26.67	35.82	0.00	0.00	0.00	66.67	20.69	31.58

4 Effect on Pronominal Coreference Resolution

The automatic disambiguation of the singular neuter pronoun finds it motivation in the difficulty to handle these cases in automatic coreference resolution. Our focus is on a classification based approach to coreference resolution, as for

example described by [8], [9], [10], [6] and others, in which information about potentially coreferring pairs of NPs is represented as a set of feature vectors which are then classified by a machine learning algorithm as being coreferential or not. Instances are created between every NP and all of its preceding NPs. Sometimes, the search scope is limited through the application of distance restrictions or linguistically motivated filters (see for example [11–13]). Applied to the case of the Dutch pronominal "het", this implies that for each occurrence of the pronoun, an instance is created. The automatic detection of the non-referential uses of the pronoun could lead to the creation of instances solely for the occurrences of "het" which do refer to a preceding noun phrase.

In a postprocessing phase, a complete coreference chain is built between the pairs of NPs that were classified as being coreferential. If we consider the task of pronominal coreference resolution, two types of errors can occur on the coreference chain level, namely precision and recall errors. In a coreferential chain, all discourse entities (mostly noun phrases) referring to each other are gathered in one single chain. The recall errors are caused by classifying positive instances as being negative. These false negatives cause missing links in the coreferential chains, as exemplified in (6) and (7), in which the pronoun was classified as being not coreferential with any of the preceding NP's.

(6)   The company will work with Sega Enterprises of Japan, SegaSoft and Time Warner Interactive to test the software. **It** will be sold starting this summer. (MUC-7)

(7)   Maar voorzitter Spiritus-Dassesse gelooft niet in het nieuwe plan. **Het** lijkt te veel op het vorige. (KNACK)
English: But chairwoman Spiritus-Dassesse does not have faith in the new plan. **It** resembles the previous one too much.

The precision errors on the other hand are caused by classifying negative instances as being positive and create spurious links in the coreference chains, as shown in (8), in which the pronoun is erroneously linked to "the US government" and in (9), in which an antecedent is sought for the non-referential "het".

(8)   **Hughes Electronics Corp.** has paid *the U.S. government* $4 million to settle a 1990 lawsuit filed by two former employees who accused **it** of lying to the Pentagon. (MUC-7)

(9)   Een god van *het vuur*. Als vice-minster van Defensie heeft Paul Wolfowitz eigenlijk een bescheiden job in de Amerikaanse regering. Hoe komt **het** dan dat hij zoveel invloed heeft in het Witte Huis? (KNACK)
English: A god of *the fire*. As a vice minister of Defense, Paul Wolfowitz has a rather insignificant job in the American government. How is **it** possible that he has such an influence in the White House?

In order to evaluate the effect of this classification of "het" on pronominal coreference resolution for Dutch, we performed a 10-fold cross-validation experiment using TIMBL on the 242 annotated Knack documents, which were also annotated with coreferential chain information. The search scope for instance

construction was reduced to 3 sentences and the instances consist of a set of features encoding morphological-lexical, syntactic, semantic, string matching and positional information sources [6].

The following experiments were conducted. In a first experiment, the output of the experiments described in Section 3 was used as the basis for filtering (Predicted). This implies that only the instances of "het" which were classified as referring to a preceding NP, were taken into account for coreference resolution. However, given the low F-score on this category in the Knack corpus, we performed a second experiment in order to assess the upper bound of potential performance increase. For this experiment we used the annotated corpus as an oracle to filter out all NPs not referring to a preceding noun phrase (Oracle). Table 3 shows the classification results before and after filtering on the instances in which "het" occurs as a potential anaphor. It reveals that filtering leads to a large reduction of the instances, but to a decrease in F-score. Furthermore, the expected potential performance increase is low (2%).

Table 3. Classification performance on the instances in which "het" occurs as potential anaphor. These are the results before and after filtering.

	#number	accuracy	precision	recall	F-score
Default	9322	97.71	11.58	7.86	9.36
Oracle	1719	90.98	19.23	8.13	11.43
Predicted	2604	95.20	7.69	4.94	6.02

The low classification results show that filtering is insufficient to detect the correct antecedent for a given anaphor. In addition to filtering, more effort should be put in new features on top of the current 39 morphological-lexical, syntactic, semantic, string matching and positional features in order to detect the appropriate referent for an anaphoric "het". Consider for example the two instances in (10), which contain too little evidence to decide on a positive or negative classification.

(10) (het) (aids) 1 5 heeft , aangezien WW(pv,tgw,met-t) LET() VG(onder) een klasse van LID(onbep,stan,agr) N(soort,ev,basis,zijd,stan) VZ(init) dist_lt_two appo_no jpron_yes 0 0 0 def_yes num_yes 0 0 0 0 0 0 0 0 I-OBJ 0 0 object 0 0 0 0 0 POS
(het) (Het ogenschijnlijke doel) 19 155 heeft , aangezien WW(pv,tgw,met-t) LET() VG(onder) een klasse van LID(onbep,stan,agr) N(soort,ev,basis,zijd,stan) VZ(init) dist_gt_two appo_no jpron_yes 0 0 0 def_yes num_yes 0 0 0 0 0 0 0 0 I-OBJ I-SU 0 object GEN_NEUT 0 0 0 0 NEG

5 Concluding Remarks

We have shown that in a classification-based machine learning approach to coreference resolution for Dutch, the accurate disambiguation of "het" (it) as being

referential or not can lead to modest improved performance on the resolution of pronominal coreference. We developed a machine learning based system for the disambiguation of referential or non-referential use of "het" using memory-based learning and genetic algorithm based joined optimization of feature selection and algorithm parameter selection. Results show that filtering is a first step towards an improved pronominal resolution and that the selection of the appropriate referent for an anaphoric "het" remains problematic. In addition to filtering, more effort should be invested in discriminating features capturing the relationship between an anaphoric "het" and its referent.

References

1. Paice, C., Husk, G.: Towards an automatic recognition of anaphoric features in english text: the impersonal pronoun 'it'. Computer Speech and Language **2** (1987) 109–132
2. Lappin, S., Leass, H.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994) 535–561
3. Boyd, A., Gegg-Harrison, W., Byron, D.: Identifiying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In: Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP. (2005) 40–47
4. Evans, R.: Applying machine learning toward an automatic classification of it. Literary and Linguistic Computing **16**(1) (2001) 45–57
5. Ng, V., Cardie, C.: Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002). (2002)
6. Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, Antwerp University (2005)
7. Daelemans, W., van den Bosch, A.: Memory-based Language Processing. Cambridge University Press (2005)
8. McCarthy, J.: A Trainable Approach to Coreference Resolution for Information Extraction. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst MA (1996)
9. Soon, W., Ng, H., Lim, D.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27**(4) (2001) 521–544
10. Ng, V., Cardie, C.: Combining sample selection and error-driven pruning for machine learning of coreference rules. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). (2002) 55–62
11. Yang, X., Zhou, G., Su, S., Tan, C.: Coreference resolution using competition learning approach. In: Proceedings of the 41th Annual Meeting of the Association for Computatiational Linguistics (ACL-03). (2003) 176–183
12. Uryupina, O.: Linguistically motivated sample selection for coreference resolution. In: Proceedings of DAARC-2004. (2004)
13. Hendrickx, I., Hoste, V., Daelemans, W.: Evaluating hybrid versus data-driven coreference resolution. In: Anaphora: Analysis, Algorithms and Applications (LNAI 4410). (2007)