Improving an Anaphora Resolution System for Norwegian

Christer JOHANSSON
Linguistics, Literary and Estethical Studies
Section for Computational Linguistics
University of Bergen, Norway
christer.johansson@uib.no

Anders NØKLESTAD
The Text Laboratory
University of Oslo
anders.noklestad@iln.uio.no

1 Introduction

Anaphora Resolution has proven to be a formidable task for computers to perform, and yet humans can easily, it seems, follow the reference trails in written or spoken discourse. Some of the human ability stems from syntactic restrictions. Although such restrictions may work well within sentences, it proves harder to follow the trail across sentence boundaries. And even within sentences it is not always clear which antecedent to choose. In example 1a it is not entirely obvious whose car is to be washed, the washer's car or the teller's car. In many such cases it is clear from the situation, which is in turn recreated in the mind of the reader from a textual discourse. There are also cases such as 1b and 1c that are in theory completely unambiguous, but in practice sometimes confused with each other.

(1) a. He_1 told him_2 to wash his car.
b. He_1 told him_2 to wash himself_2
c. He_1 told him_2 to wash him_1

The second problem of anaphora resolution is that referring nouns may actually refer to many different entities in the world. A mention of 'Ford' may refer to a shallow part of a river, a name of somebody named Ford, who might be the founder of the Ford company, the present leader of the Ford company, the ex-president Gerald Ford, or the guitarist and blues singer Robben Ford. Ford may also be a brand name, or an actual Ford car. The Ford Company might be referred to by *it*, or by *they*, and the car may be referred to by *it*, or affectionately as a *he* or *she*. There are in fact many ways to weave a net of possible threads through a discourse, and yet there are clearly ways that make a text be understandable, and other ways that obscure and mislead the reader.

We have introduced a few reasons why we should not expect a perfect result from an anaphora resolution algorithm. The first part of the argument is that there are cases that are genuinely hard for humans to properly resolve. Humans may overcome this by two different means: 1) the writer or speaker may structure the discourse so that it minimizes the inferential load of the recipients. 2) The recipient may employ world knowledge and infer the most likely intended content. Both these alternatives involve having a model of the mind of the other. If the sentence producers do structure the discourse so that it will be clearer, then we might have a chance to induce this structure from data on usage. However, there is a good chance that there is considerable variation in how well a writer structures discourse, and what a recipient may tolerate. Some of this variation may originate in slight differences in experience and world knowledge. Let us now go on to look at some previous research and our results for Norwegian.

1.1 Previous research

In much of the past research, the focus has been very much on how syntactic structures block or allow coreference. One early approach (Hobbs, 1978) relies on having parsed syntactic trees available. Another approach that assumes more shallow knowledge is to model the salience of possible antecedents through the use of weighed salience factors (Lappin and Leass, 1994) or indicators (Mitkov, 2002).

Centering Theory (Grosz et al., 1995) has been influential for the development of Anaphora Resolution systems. Centering assumes a salience rating as well, but has an added constraint that there is a single item which is in focus at any given time and therefore most likely to be pronominalized.

Recently, anaphora resolution has been performed by machine learning techniques, using resources that are less demanding. Cardie and Wagstaff (1999) introduced an unsuper-

Johansson, C. (Ed.)
Proceedings of the Second Workshop
on Anaphora Resolution (2008)

vised clustering technique. Coreference can also be viewed as a classification task. A comparison between decision tree learning (classification) and the clustering algorithm, shows, not surprisingly, that training on pre-classified examples can provide better results (Soon et al., 2001).

2 The BREDT system

We have developed a system for resolution of the Norwegian pronouns *han* "he", *hun* "she", *seg* "himself/herself/itself/themselves", *den* "it" (for masc./fem.), and *de* "they". We decided to work with memory-based learning, as implemented in the TiMBL software package (Daelemans et al., 2004). Our system is described elsewhere (Johansson and Nøklestad, 2005; Johansson et al., 2006; Nøklestad, forthcoming), but briefly the system uses stored classified exemplars of generalized possible anaphor/antecedent pairs. Some of the lexical context, sentence distance, morphological information and syntactic roles of both antecedent candidate and anaphor are abstracted into features and match-features (same value on the feature or not). This information is stored in an instance with a class denoting the coreference of the pair. For the collection of such pairs, we calculate the information gain for each feature (i.e. how much does each feature contribute to a correct decision on coreference).

We have tested many feature combinations to find an optimal set for the task. Recent results using the features listed in Table 1 (Nøklestad, forthcoming) show an overall accuracy of 73.53%, earlier results were about 11 percentage points lower (Johansson et al., 2006).

	Accuracy
Identical anaphor/antecedent	89.35%
Non-ident. anaphor/antecedent	54.15%
Overall	73.53%

2.1 Anaphora Resolution Corpus

The latest version of the corpus consists of 25,451 (word) tokens, 1778 of which are pronouns that are included in the systems capabilities. From these 1778 anaphor candidates, 1701 (95.7%) are marked with a coreferential antecedent in the text. The annotation and the annotation guidelines for our corpus are now available (Borthen et al., 2007). The corpus was split equally into a training section and a test section.

Anaphors	total	with referents
Training set	903	858 (95.0%)
Test set	875	843 (96.3%)
Total	1778	1701 (95.7%)

2.2 Ongoing and future improvements

We have found one point where the automatic coreference finding could be more easily improved, namely the cases where antecedents and anaphors are different.

From table 1 we see that the strongest features are those that involve match between lemma forms. The second strongest is what we call concatenated lemmas, which is a match on the concatenated antecedent and anaphor. The concatenated lemma could provide some information on which lemmas tend to corefer. Since this feature is so strong, it is interesting to try to generalize this further using some other strategy than just storing them from the training corpus.

Improvements may come from semantic information which may be acquired by finding default pronouns for names, finding occupations (journalist, writer, police) that are likely to take a human pronoun, and finding things and part-of relations. Recent research (Markert and Nissim, 2005; Markert et al., 2003, inter al.) has shown that the web provides very good coverage, and high precision is obtained, despite the lack of annotation, because the size of the corpus allows us to select and use patterns that are very precise. The results compare well to results obtained from using very large corpora with extensive mark-up, such as the British National Corpus. Another source of information are gazetteers, available on the web via various government organizations.

We are currently using the web to extract values for the "In animate noun list" and "Online animacy evaluation" features in Table 1. For these features, we formulate a set of linguistic constructions which typically involve animate nouns, and we submit these patterns to the Google search engine in order to harvest as many animate nouns as possible (for the first feature) or evaluate the animacy of a given antecedent candidate (for the second one). More details are provided by Nøklestad (forthcoming). We are also using gazetteers as one of the information sources for the the named en-

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129

Feature	han/hun	den	de
Lemma match	34.70	23.71	26.29
Reflexive and closest subject	28.33	0	0
Non-reflexive and subject of same sent.	1.35	1.08	0.13
Subject antecedent	9.64	0.03	2.90
Syntactic parallelism	2.23	1.28	1.18
Gender match	5.05	11.39	0
Distance < 2 sentences	1.42	1.78	2.56
Distance < 3 sentences	1.18	0.46	1.83
Concatenated lemmas	8.70	15.38	6.90
In animate noun list	10.06	16.82	0.99
Online animacy evaluation	2.21	0.12	1.29
Antecedent is a person	2.17	1.32	0.84

Table 1: Optimal Features with Gain Ratio Weights*100

tity recognition module which provides values for the "Antecedent is a person" feature.

We will continue to investigate how we can use the web to gain more detailed and useful information with a higher degree of coverage than is available from manually compiled ontologies, a knowledge source already proven useful for anaphora resolution (Lech and deSmedt, 2007). For example, at the moment we have work in progress aimed at named entity recognition using the web (Rømcke, in progress). However, web searches for linguistic purposes are still limited by a lack of useful functions such as the ability to use linguistic wild cards. Such wild cards could specify a linguistic lexical category, for instance a tensed verb, or a syntactic category such as an entire noun phrase. Wild cards are available in some of the search interfaces that are specifically designed for linguistic research (e.g. the one for the Oslo Corpus of Tagged Norwegian Texts[1]), but are currently not available for web searches.

References

K. Borthen, L. Johnsen, and C. Johansson. 2007. Coding anaphor-antecedent relations – the annotation manual for bredt. In C. Johansson, editor, *Proceedings from the first Bergen Workshop on Anaphora Resolution (WAR I)*, pages 86–111, Cambridge, UK. Cambridge Scholars Publishing.

C. Cardie and K. Wagstaff. 1999. Noun phrase coreference as clustering. In *Proc. of the joint SIGDAT Conf. on Empirical Methods in NLP and Very Large Corpora*, pages 82–89.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van der Bosch. 2004. TiMBL: Tilburg Memory-Based Learner, Version 5.1, Reference Guide. Technical Report ILK 04–02, the ILK Group, Tilburg, the Netherlands.

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

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

C. Johansson and A. Nøklestad. 2005. Detecting Reference Chains in Norwegian. In *Proceedings of the 15th NoDaLiDa Conference*, pages 1–10, Joensuu, Finland. University of Joensuu electronic publications in linguistics and language technology.

C. Johansson, A. Nøklestad, and Ø. Reigem. 2006. Developing a re-usable web-demonstrator for automatic anaphora resolution with support for manual editing of coreference chains. In *Proceedings of LREC 2006 - 5th International Conference on Language Resources and Evaluation*, pages 1161–1166, Paris. European Language Resources Association.

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

T. Lech and K. deSmedt. 2007. Ontology extraction for coreference chaining. In C. Johansson, editor, *Proceedings from the first Bergen Workshop on Anaphora Resolution (WAR I)*, pages 26–38, Cambridge, UK. Cambridge Scholars Publishing.

K. Markert and M. Nissim. 2005. Comparing knowledge sources for nominal anaphora res-

[1] http://www.tekstlab.uio.no/norsk/bokmaal/english.html.

olution. *Computational Linguistics*, 31.

K. Markert, N. Modjeska, and M. Nissim. 2003. Using the web for nominal anaphora resolution. In *EACL Workshop on the Computational Treatment of Anaphora*, citeseer.ifi.unizh.ch/markert03using.html.

R. Mitkov. 2002. *Anaphora Resolution*. Longman/Pearson Education, London, UK.

A. Nøklestad. forthcoming. *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment, Disambiguation, and Animacy Detection*. University of Oslo, Ph.D. thesis.

A. Rømcke. in progress. *Named Entity Recognition using the Web: the Truth is out there*. University of Bergen, dept. of Linguistics.

Wee Meng Soon, Hwee Tou Ng, and D. Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

ISSN 1736-6305 Vol. 2
http://dspace.utlib.ee/dspace
/handle/10062/7129