Data-driven Coreference Resolution for Swedish

Fredrik Axelsson*, Birger Rydback*, Fredrik Johansson*, Jonatan Bengtsson†, Svetoslav Marinov†

*Chalmers University of Technology, †Findwise AB
Gothenburg
fredrik.axelsson@chs.chalmers.se, rydback@gmail.com, frejohk@chalmers.se
†{jonatan.bengtsson, svetoslav.marinov}@findwise.com

Abstract

In this paper we propose a new Coreference Resolution system for Swedish, based on supervised machine learning methods trained on the SUC-core dataset. Our method improves on state-of-the-art results for the data, achieving an average F_1-score of 50.9 using the standard CoNLL 2012 metrics.

1. Introduction

Two main research paradigms have gained prominence in the domain of Coreference Resolution (CR) - Knowledge-based and Data-driven methods. Top knowledge-based systems (Raghunathan et al., 2010) employ large sets of linguistic rules to deterministically classify pairs of mentions. While easy to analyze linguistically, such systems are difficult to adapt to new languages and domains. On the other hand, data-driven systems require access to annotated data. As the data-driven approaches have successfully been applied to a number of Natural Language Processing (NLP) tasks, the availability of corpora marked with coreference information has made them favourable candidates. In previous research, the focus has been on languages such as English, Chinese and Arabic, included in the Conference on Computational Natural Language Learning (CoNLL) datasets (Pradhan et al., 2012), while Swedish has received little attention.

Earlier works in Swedish CR include a data-driven approach by Nilsson (2010), a knowledge-based implementation by Byström (2012) along with an annotated corpus. The corpus consists of the core part of the balanced Swedish corpus Stockholm-Umeå Corpus (SUC) along with coreference information (Nilsson Björkenstam, 2013). While Nilsson annotated an unreleased corpus for her work, Byström used a beta version of SUC to evaluate his work and is until now the only work to present results on this data.

In this paper we present a new CR system for Swedish, based on supervised machine learning methods trained on SUC-core. Our method improves on state-of-the-art results for the data set.

2. Task definition

The task of a CR system is to find all the entities (so called mentions) in a text which refer to the same real world entity. We follow the standard definition of a coreference relation which exists between pairs of mentions that may be either pronouns, noun phrases or Named Entities. Consider:

"*I* voted for *Nader* because *he* was most aligned with *my values*", *she* said.

In this sentence there are six mentions that can be divided into the following groups {*I*, *my*, *she*}, {*Nader*, *he*} and {*my values*}, where each group consists of different mentions of the same real world entity.

We concentrate only on anaphoric relations and exclude cataphora, bridging and bound anaphors. Non-referential *det* (it) and the indefinite pronoun *man* (one) are also excluded.

3. Data-driven methods

During the last decade most CR systems have been based on a two step approach (Ng, 2010), where mention pairs are first extracted and subsequently clustered. The first data-driven method was an approach using the C5 decision-tree learning algorithm (Soon et al., 2001). Fernandes et al. (2012) presented the best performing system at CoNLL 2012 shared task, based on Latent Structured Perceptron (LSP). Their system was later improved by Björkelund and Kuhn (2014). One of the biggest advantages of LSP for NLP tasks is its ability to take structural information into consideration when making predictions (Collins, 2002). We will therefore rely on the LSP for the current task.

4. System Architecture

Our system consists of three main stages: mention detection, pair selection and finally clustering. In the first step, the mentions are extracted from the training data and the head word of each mention is identified as well as some additional information about the mentions. After selecting the pairs a feature vector is generated for each pair, see Section 4.2. Finally, the core model clusters the mentions given the features. In a post-processing step all singletons are removed.

4.1 The core algorithms

We implement two different algorithms. One is based on a decision tree generator, C4.5 (McCallum, 2002). During prediction, this algorithm classifies mention pairs as referential based on a confidence score over the threshold 0.5, the pairs are then clustered using Aggressive Merge Clustering (McCarthy and Lehnert, 1995).

The second is based on LSP à la Fernandes et al. (2012) and Björkelund and Kuhn (2014). We follow the implementation outlined in Fernandes et al. (2012), using their definition of the loss function and procedure for updating the perceptron weight vector. However, instead of using Maximum Branching Tree to find an optimal scoring tree

we rely on Best First Search used by Björkelund and Kuhn (2014).

4.2 Features

A Fundamental part of the data-driven methods is the selection of features used during the learning and decision. We are striving to identify features which model the similarity of mentions within a cluster (class) but increase the differences among the clusters (classes).

The feature set in the current experiment is a combination of features used by Nilsson (2010) and Fernandes et al. (2012). However, we only chose those which reflect the information present in SUC-core and the linguistic properties of Swedish. The SUC-core provides annotations for PoS-tags, morphological information, token lemmas, compound word splitting, semantic information for named entities and mentions.

By translating the lemma of the nouns, the English WordNet is used in order to find additional semantic information. Semantic information for named entities, such as locations, organizations and people, is annotated in the corpus. For people, gender is assigned using lists of male and female names. Whenever applicable, pronouns are explicitly enriched with animacy, number and gender.

Lexical	Lemma and string comparison, string overlap
Syntactic	Named entity match, number match, subject and object match, inside quotes, pronoun info match, demonstrative match, definiteness match, common gender match, PoS tag and PoS tag match
Semantic	Semantic type match, animacy match, gender match
Distance	Mention and sentence distances, nested mentions

Table 1: Features from four categories are used. For more details see Axelsson and Rydback (2014).

In addition to homogeneous features we consider combination of features. These are constructed from the decision tree generated by the C4.5 algorithm and called feature templates, in the same fashion as Fernandes et al. (2012), see Axelsson and Rydback (2014) for more details.

5. Experimental Setup

Three models were trained on SUC-core, one C4.5 and two LSP (with and without feature templates). During both training and testing, the SUC-core annotated mentions were used and mention pairs which had a distance greater than a given threshold in between were not considered. Other than considering distance, no additional filter was applied. Initial experiments showed good performance with a distance threshold of 160 mentions.

Finally, in order to use the standard CoNLL evaluation metrics, the output was converted to fit the required format.

Algorithm	**MUC**	**B^3**	**$CEAF_e$**	**Avg**
LSP with Templates	65.7	**45.5**	**41.6**	**50.9**
LSP w/o Templates	65.2	43.0	40.9	49.7
C4.5	**68.6**	45.1	38.7	50.7
Byström (2012)†	≈30	–	–	–
Nilsson (2010) ††	67.4	–	–	–

Table 2: Top results for Swedish CR. † The knowledge-based approach by Byström (2012) was evaluated on a beta version of the SUC-core corpus. †† The hybrid approach by Nilsson (2010) was evaluated on a different corpus.

6. Results

Table 2 shows the results (F_1-score) of our system. These are calculated as a corpus average with leave-one-out cross validation, using the scoring strategy and scorer (version 7) from the 2012 CoNLL shared task (Pradhan et al., 2012).

Our current models give the best results on the only publicly available Swedish corpus with CR information, SUC-core, for all metrics, with the LSP with templates yielding the best overall performance. While not directly comparable, as it is evaluated on a different corpus, we include the results achieved by the hybrid approach of Nilsson (2010).

7. Conclusion

In this paper we presented a data-driven method for CR for Swedish. Our system outperforms the current rule-based approach on SUC-core and fares similarly to the state-of-the-art for Swedish. While the training corpus is very small (only 2.5% of the size of the standard dataset for English) the initial results are promising.

8. Future work

By making more exhaustive research on parameter and feature combinations, the performance of the system might be improved. Also the use of a Swedish WordNet would remove errors due to ambiguities in translation. Along with this, the effects of different filtering strategies for selecting training data would be interesting to observe.

In order to improve the results further and avoid problems caused by the small dataset (such as overfitting) more annotated data is required.

Since coreference resolution is considered a domain-dependent problem in the sense that it is affected by a high diversity in genres and domains, it would be reasonable to conclude that the number of different types of documents in SUC-core affects the performance in addition to the small corpus size. Training the system on a single domain may therefore enhance the performance in the given domain.

Another path of research could be semi-supervised methods, which would fit better given the small but annotated dataset SUC-core along with the larger SUC, which lack CR information.

References

F. Axelsson and B. Rydback. 2014. Data-driven Coreference Resolution for Swedish. Master's thesis, Chalmers University of Technology.

A. Björkelund and J. Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 47–57.

E. Byström. 2012. Knowledge-based Coreference Resolution in Swedish. Bachelor's thesis, Stockholm University.

M. Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 1–8.

E. R. Fernandes, C. N. dos Santos, and R. L. Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*, pages 41–48.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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

V Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411.

K. Nilsson Björkenstam. 2013. Suc-core: A balanced corpus annotated with noun phrase coreference. In *Northern European Journal of Language Technology*, pages 19–39.

K. Nilsson. 2010. Hybrid methods for coreference resolution in Swedish. Ph.D. thesis, Stockholm University.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*.

K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and Ch. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501.

W. M. Soon, H. T. Ng, and D. Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544.