Bachelorthesis

Improving Anaphora Resolution Through Corpus Mined Gender Information

Jan Henry van der Vegte

July 2016 – October 2016

Matriculation Id: 3008277
Course of Study: Applied Cognitive and Media Science

Reviewer:

Professor Dr.-Ing. Torsten Zesch
Prof. Dr. rer. soc. Heinz Ulrich Hoppe

University of Duisburg-Essen
Faculty of Engineering
Department of Computer and Cognitive Sciences
Language Technology Lab 47057 Duisburg

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere weiterhin, dass ich diese Arbeit noch keinem anderen Prüfungsgremium vorgelegt habe.

Duisburg, im November 1492

.....

Jan Henry van der Vegte

Contents

Chapter 1

Introduction

1.1 Background

In the last decades, the amount of textual information in media has increased severely, making automatic text comprehension indispensable. Since textual data found online is mostly unstructured, which means that there is no formal structure in pre-defined manner, various information need to be added in order to make automatic understanding possible. For several natural language processing (NLP) tasks, referential relationships between words in a document need to be set.

The procedure of determining whether two expressions refer to each other, meaning that they are instances of the same entity, is called anaphora resolution. The word to be resolved is termed anaphora, while its predecessor is the antecedent. It differs from coreference resolution by only resolving words, which can only be interpreted through its antecedent (Recasens et al. 2007) (1), while all corefering expressions are considered in coreference resolution (2).

(1) [Aberfoyle] describes [itself] as [The Gateway to [the Trossachs]].
(resolve "itself" to "Aberfoyle")

(2) As late as 1790, all the residents in the parish of [Aberfoyle] spoke [Scottish Gaelic]. From 1882 [the village] was served by [Aberfoyle railway station].
(resolve "the village" to "Aberfoyle")

Resolving noun phrases is a growing task in Natural Language Processing (NLP) and increased its relevance in the last decades, that it has even developed into a standalone subtask in the DARPA Message Understanding Conference in 1995 (MUC-6 1995). The International Workshop on Semantic Evaluation (SemEval) ran a coreference resolution task on multiple languages (Recasens et al. 2010), emphasizing the importance of coreference resolution systems. There are several important applications of coreference and anaphora resolution, such as Information Extraction (IE) (McCarthy & Lehnert 1995), Question Answering (QA) (Morton 2000), and Summarization (Steinberger et al. 2007). Information Extraction has set itself the objective of summarizing relevant information

from documents. Anaphora resolution is needed, because the sought entity is often referenced through different words (for instance personal pronouns). McCarthy and Lehnert described it as a classification problem: "given two references, do they refer to the same object or different objects."
The Question Answering task described by Morton has the goal to find a 250 byte string excerpt out of a number of documents as the answer to a query. Annotated coreference chains were used to link all instances of the same entity in a document. Occurrences in an other sentence are given a lower weight for prediction. The use of annotated coreference chains improved the prediction slightly.
Steinberger et al. figured out that the additional use of anaphoric information improved their performance score over solely Latent Semantic Analysis (LSA) summarization.

A lot of different information sources including syntactic, semantic, and pragmatic knowledge is needed since selecting a possible antecedent is a decision under high ambiguity. The decisive factor for determination might be for instance gender agreement or the distance between antecedent and anaphora. Sometimes there is no decisive factor at all. Examples for the importance of gender agreement are shown in (3) and (4), the influence of word distance could simplified be described as it is more likely to find the antecedent in proximity to its anaphora.

(3) John and Jill had a date, but he didn't come. (resolve "he" to "John").

(4) John and Jill had a date, but she didn't come. (resolve "she" to "Jill").

!! Earlier anaphora resolution systems often used rule based techniques to determine the correct antecedent !!(Mitkov, 1998), but lack due to their limited generalisability. In particular, rule based techniques often require specific domain knowledge, which makes the development complex and time consuming. Furthermore, the assignment to other languages is hampered by rules.

However, machine learning could overcome these issues. !! In machine learning, the system adapts and learns general principles from experience through training data. If domain specific data should be predicted, the algorithm should be trained on that domain.

1.2 Motivation

Significant factors of uncertainty are gender and number, because they are hard to determine. At first, information is needed whether a noun is male, female, neutral, or plural. Honorifics like "Mr." and "Mrs." are gender indicators, but not sufficient due to their sparsity. Stereotypical occupations and gender indicating suffixes like policeman and policewoman turned out to be no longer reliable (Evans & Orasan 2000). For that reason, gender and number information needs to be learned from an external source.

There are two different strategies for implementing reliable gender information:
Firstly, gender can be treated as hard constraint. This means, that either the most likely gender is assigned or, in case of uncertainty, no assignment is made at all. The leading coreference resolution systems mostly use hard constraint gender information and yield accuracy similar to non-learning approaches (Soon et al. 2001). Soon et al. obtained their gender information through WordNet. The gender of to the most frequent sense of a noun is assumed.
Secondly, gender can be expressed through probabilities. If a noun is male in 70 of 100 cases, the probability for it to be male is 70 % (note that this is simplified - the distribution will be smoothed to avoid 0-probabilities). In 2005, Bergsma obtained encouraging results with the use of gender probabilities. More precisely, adding corpus mined gender frequencies improved their accuracy by approximately 10 %.

In this work i will present a machine learning approach to anaphora resolution, focusing on third-person pronominal anaphors. The two main purposes are to determine the impact of gender probability and to compare it to gender information treated as hard constraint. First of all, it should be evaluated whether the improvement through gender frequencies can be replicated on different data sets. In a second step, the gender frequencies will be replaced by the assignment of the most frequent gender to examine the influence of nothing but the gender implementation strategy. This is necessary, as usage of different data sets and algorithms makes the comparison of papers inconclusive. Finally, it needs to be examined whether the hypothesis, that corpus based gender frequencies have a higher impact than gender constraints, can be confirmed.

—-
establish your territory (say what the topic is about) and/or niche (show why there needs to be further research on your topic) shortly introduce your research/what you will do in your thesis (make hypotheses; state the research questions)

Chapter 2

Related Work

Anaphora resolution systems emerged into two different strategies. The first one are rule-based techniques, which focus more on theoretical considerations. The second strategy uses machine learning and is based on annotated data. In the following chapter i will briefly present both and discuss their advantages and disadvantages.

2.1 Rule-based techniques

Rule-based techniques rely on human understandment of syntactic and semantic principles of natural language. Clues that could be helpful for identifying the antecedent are manually implemented as rules. To identify relevant clues, prior knowledge about linguistic principles (such as binding principles) is necessary. Since rules might be domain-specific, the implementation would most likely be worse on other domains. Refinements for different domains would make the development even more complex and time-consuming.

2.1.1 Knowledge-poor anaphora resolution

A domain independent approach by (Mitkov 1998) tried to eliminate the disadvantages of previous rule-based systems. Mitkov renounced complex syntax and semantic analysis in order to keep the algorithm as less domain specific as possible. The algorithm was informally desribed by Mitkov in three steps:

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases only to the left of the anaphor
2. Select from the noun phrases identified only those which agree in gender and number with the pronominal anaphor and group them as a set of potential candidates
3. Apply the antecedent indicators to each potential candidate and assign scores; the candidate with the highest aggregate score is proposed as antecedent

Antecedent indicators are features with a score of -1,0,1, or 2 and are for instance informations like Definiteness (whether the noun phrase contains a definite article) or

provide background information needed to understand your thesis assures your readers that you are familiar with the important research that has been carried out in your area establishes your research w.r.t. research in your field

2.2 Machine learning-based techniques

e.g.

- conceptual framework
- structured overview on comparable approaches
- different perspectives on your topic

a**?** ad

Chapter 3

Concept

Chapter 4

Implementation

Chapter 5

Evaluation

Chapter 6

Conclusion

6.1 Summary

What was done? What was learnt?

6.2 Outlook

What can/has to be/may be done in future research? Impact on other branches of science? society?

Appendix

References

List of Figures

List of Tables

Listings

References

Evans, R., & Orasan, C. (2000). Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the discourse anaphora and reference resolution conference (daarc2000)* (pp. 154–162).

McCarthy, J. F., & Lehnert, W. G. (1995). Using decision trees for coreference resolution. *arXiv preprint cmp-lg/9505043*.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics-volume 2* (pp. 869–875).

Morton, T. S. (2000). Coreference for nlp applications. In *Proceedings of the 38th annual meeting on association for computational linguistics* (pp. 173–180).

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., ... Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 1–8).

Recasens, M., Marti, M. A., & Taulé, M. (2007). Where anaphora and coreference meet. annotation in the spanish cess-ece corpus. In *Proceedings of ranlp*.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, *27*(4), 521–544.

Steinberger, J., Poesio, M., Kabadjov, M. A., & Ježek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing & Management*, *43*(6), 1663–1680.