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 and 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 name is male or female. Honorifics like "Mr." and "Mrs." are gender indicators, but not sufficient due to their sparsity. Stereotypical occupations and gender indicating suffixes turned out to be no longer reliable (Evans and Orasan, 2000). "SINGULAR PLURAL WORDS?" For that reason,

gender information needs to be learned In this work i will present a machine learning approach to anaphora resolution, focusing on third-person pronominal anaphors.

In 2005, Bergsma presented a machine learning approach for anaphora resolution, that automatically In this work i will present a machine learning approach to anaphora resolution for third-person pronominal anaphors

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

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

e.g.

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

?

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