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 even became a standalone subtask in the DARPA Message Understanding Conference in 1995 (MUC-6 1995). The International Workshop on Semantic Evaluation (SemEval) conducted a coreference resolution task on multiple languages (Recasens et al. 2010) emphasizing its importance. There are several fundamental 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 targets to summarize relevant information from documents.

Anaphora resolution is required, as the quested entity is often referenced through various words !!(for instance personal pronouns). (McCarthy & Lehnert 1995) described the latter as a classification problem: "Given two references, do they refer to the same object or different objects."
The question answering task described by Morton seeks 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 another 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.

Various information sources including syntactic, semantic, and pragmatic knowledge are needed since selecting a possible antecedent is a decision under high ambiguity. The decisive factor for determination might be e.g. 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").

-> PRONOUN RESOLUTION: A.I. complete' task– (CogNIAC)

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 (Soon et al. 2001). 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 %.

This work 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. First of all, there are rule-based techniques which focus more on theoretical considerations. The second strategy uses machine learning and is based on annotated data. The following chapter will briefly present both and discuss their advantages and disadvantages, followed by exemplary realisations.

2.1 Rule-Based Techniques

Rule-based techniques rely on manual understanding and implementation of syntactic and semantic principles in natural language (Kennedy & Boguraev 1996; Mitkov 1994). Clues that could be helpful for antecedent identification 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. Nevertheless, rule-based techniques are much more transparent in contrast to machine learning.

2.1.1 Anaphora Resolution with Limited Knowledge

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 described 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 for instance definiteness (whether the noun phrase contains a definite article) or non-prepositional noun phrases (whether the noun phrase is part of a prepositional phrase). A positive indicator score increases the likelihood that the selected noun phrase is the antecedent and vice versa.

2.1.2 CogNIAC

Another rule-based approach was presented by (Baldwin 1997) with CogNIAC, a high precision pronoun resolution system. It only resolves pronouns when high confidence rules (shown in Table 2.1) are satisfied in order to avoid decisions under ambiguity and to ensure that only very likely antecedents are attached (high precision). This might lead to a neglect of less likely but still correct antecedents and lower the recall score.
For each pronoun the rules are applied one by one. If the given rule has found a matching candidate it will be accepted. Otherwise the next rule will be applied. If none matches the candidates it will be left unresolved as this implicates a higher ambiguity. In order to apply Baldwins high confidence rules, information on sentences, part-of-speech, and noun phrases is required and therefore annotated. Semantic category information such as gender and number is determined through various databases.
As can be seen the order of rules lead from higher to lower precision: if only one possible antecedent can be found (rule 1) it is most likely the correct antecedent while rule 6 indicates more ambiguity as it relies on more content-related information. Human understanding of syntax and semantics is needed to determine a specific order of rules. Therefore, adding new rules might not improve the performance even though those rules are reasonable in itself. Most rule-based systems struggle with that problem.

Table 2.1: CogNIAC core rules

Rule	Description
1) Unique in Discourse	If there is a single possible antecedent PAi in the read-in portion of the entire discourse, then pick PAi as the antecedent.
2) Reflexive	Pick nearest possible antecedent in read-in portion of current sentence if the anaphor is a reflexive pronoun
3) Unique in Current + Prior	If there is a single possible antecedent i in the prior sentence and the read-in portion of the current sentence, then pick i as the antecedent:
4) Possessive Pro	If the anaphor is a possessive pronoun and there is a single exact string match i of the possessive in the prior sentence, then pick i as the antecedent:
5) Unique Current Sentence	If there is a single possible antecedent in the read-in portion of the current sentence, then pick i as the antecedent
6) Unique Subject/ Subject Pronoun	If the subject of the prior sentence contains a single possible antecedent i, and the anaphor is the subject of the current sentence, then pick i as the antecedent

2.1.3 Hobbs algorithm

The Hobbs algorithm described by (Hobbs 1978) relies on parsed syntax trees containing the grammatical structure. Put simply, the tree containing the anaphora is searched left-to-right with breadth-first search and the algorithm stops when a matching noun phrase is found. Noun phrases mismatching in gender or number are neglected. As long as no matching antecedent is found, the preceding sentence will be searched.

2.2 Machine Learning-Based Techniques

Most machine learning-based techniques learn principles from annotated text corpora (Soon et al. 2001; Bergsma 2005), which include the correct label for each instance. In this context, a label will contain the information whether a noun phrase is the antecedent. A decisive factor of machine learning is, that irrelevant information (presented through features) has a lower impact on success factors (the accuracy for instance) compared to rule-based techniques, as the algorithm automatically learns to rate those as irrelevant and vice versa. Therefore, machine learning approaches tend to have little

information on linguistic principles as the algorithm should learn those autonomously. This causes the algorithm to be less domain specific, but increases the risk to miss relevant clues. However, top-performing machine learning approaches achieve accuracy scores comparable to state-of-the-art non-learning techniques (Soon et al. 2001).
Additionally, machine learning algorithms are usually more time-consuming due to the learning process.

2.2.1 Anaphors in Coreference Resolution

In coreference resolution all noun phrases referring to the same entity in the real world should be linked. The most common kind of storing coreferential information is through coreference chains, in which the current element always points towards the following same entity-element. Pronominal anaphors are included and can be extracted by choosing the previous entity of the same coreference chain. A coreference resolution system proposed by (Soon et al. 2001) used a set of 12 different features. It covers, inter alia, a distance feature (standing for the distance in sentences between two elements), a gender agreement feature (whether the gender matches), and a number agreement feature (whether the number matches). To specify most of the features, several preprocessing steps, such as noun phrase identification and part-of-speech tagging, need to be executed. For instance, sentence segmentation is required to determine the sentence distance. Deriving gender information of a noun requires information of their semantic classes. Soon et al. assumed that the semantic class of a noun phrase is the semantic class of the most frequent sense of the considered noun in WordNet. Gender agreement can be assumed, if both phrases got the same semantic class (for example "male") or if one is the parent of the other (such as phrase one is considered as "person" and phrase two as "male").

In order to make machine learning possible, training instances need to be generated.
To generate positive training instances Soon et al. used every noun phrase in a coreference chain and its predecessor in the same chain. Each intervening noun phrase forms a negative instance with the considered noun phrase.

(Ng & Cardie 2002) extended their work and improved it through additional features, a different training set creation, and a clustering algorithm to find the noun phrase with the highest likelihood of coreference. The majority of the new features is based on syntactical principles. For instance, binding constraints must be fulfilled and one phrase is not allowed to span another. Positive training instances are not created through their preceding antecedent, but through their most confident one. In addition, they started to search for a related antecedent from right-to-left for a highly likely antecedent (in contrast to starting the right-to-left search for the first previous noun phrase). Ng and Cardie reported a significant increase in precision and F1-measure compared to the initial approach by (Soon et al. 2001).

2.2.2 Bergsma

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

2.3 Compared Evaluation

2.4 A Statistical Approach

e.g.

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

a

Chapter 3

Data

3.1 Anaphora Corpus

3.2 Gender Corpus

Chapter 4

Methodology

4.1 Featureset

4.1.1 Pronoun Features

4.1.2 Antecedent Features

4.1.3 Pronoun-Antecedent Features

4.1.4 Gender Features

4.2 Generating Training Instances

4.3 Baseline Approach

4.4 SVM Classifier

Chapter 5

Evaluation

5.1 Learning curves

5.2 Feature Contribution

5.3 Error Analysis

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

References

Baldwin, B. (1997). Cogniac: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a workshop on operational factors in practical, robust anaphora resolution for unrestricted texts* (pp. 38–45).

Bergsma, S. (2005). Automatic acquisition of gender information for anaphora resolution. In *Conference of the canadian society for computational studies of intelligence* (pp. 342–353).

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).

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, *44*(4), 311–338.

Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of the 16th conference on computational linguistics-volume 1* (pp. 113–118).

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

Mitkov, R. (1994). An integrated model for anaphora resolution. In *Proceedings of the 15th conference on computational linguistics-volume 2* (pp. 1170–1176).

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).

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 104–111).

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.