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 (Chinchor & Sundheim 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) and Question Answering (QA) (Morton 2000).
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.

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").

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; Ingria & Stallard 1989). 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 The Naive Hobbs algorithm

The Naive 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. The algorithm also limits the list of possible antecedents, as for instance the antecedent can not occur in the same non-dividable noun phrase. As long as no matching antecedent is found, the preceding sentence will be searched successively.
Hobbs reported an accuracy score of 88,3 % on the pronouns ”he”,”she”,”it”, and ”them”

with only using the algorithmic approach. The usage of additional constraints improved the accuracy to 91.7 %.

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. Confirming their prediction, (Baldwin 1997) reported a high precision score (97 %), but lower recall (60 %) on their training data consisting of 198 pronouns.
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.
In a second evaluation, CogNIAC was compared to the Hobbs Algorithm (Baldwin 1997; Hobbs 1978) on singular third-person pronoun resolution. In order to maximize the ambiguity, the training data texts were narrations about same gender characters. To make accuracy scores comparable, Baldwin (1997) added lower precision rules, such as the most recent antecedent should be picked if no other rule found a matching noun phrase. The Accuracy scores reported were nearly equal (78,8% on the Hobbs Algorithm, 77,9% on CogNIAC), underlining the reason of existence of various approaches.

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 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. Only a part-of-speech tagger and a simple noun phrase identifitcation module were applied. 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

Overall, a set of 10 antecedent indicators were used which indicate either a high or a low likelihood for the noun phrase to be the antecedent. Negative indicators such as definiteness (whether the noun phrase contains a definite article, whereby indefinite

phrases decrease the likelihood) and positive indicators like term preference (if the noun phrase is a term in the field, the likelihood is increased). The score values are integers from -1 to 2.
Mitkov reported a success rate of 89,7 % on random sample texts of technical manuals. A modified approach could also be applied for polish (Mitkov & Stys 2000) and arabic (Mitkov et al. 1998) with similar success rates. A comparing evaluation to Baldwins CogNIAC (Baldwin 1997) indicated a superiority of Mitkovs approach (Mitkov 1998) as CogNIAC had a lower success rate of approximately 15 % on the previously described data set. The stated reason for the comparison was that the approaches showed several similarities as both require few preprocessing and gain their information mostly from part-of-speech tags and noun phrases.
The superiority of (Mitkov 1998) could be explained by its handling of uncertainty as the antecedent indicators are not implemented as hard constraints. Basically, Mitkovs anaphora resolution system can be described as a combination between rule-based and statistical techniques in order to use the best of both worlds.

In 2002, a revised version of the original approach by Mitkov was presented (Mitkov et al. 2002). The improved version of the original algorithm called MARS had some smaller and greater changes:
First of all, three new antecedent indicators and a module for identification of pleonastic pronouns[1] and non-nominal pronominal anaphors were added. Additionaly, the implementation of some previous features was changed as other preprocessing tools were used.

2.1.4 GuiTAR

With GuiTAR, a modular anaphora resolution tool was developed (Poesio & Kabadjov 2004). It was designed to be domain-unspecific and usable off-the-shelf which means that preprocessing steps such as part-of-speech tagging and named entity recognition will be added on itself. Either raw text data or XML files can be used as the input. In case of raw text data, XML files with annotated part-of-speech tags, noun phrase boundaries, pronoun categories etc. will be created. The anaphora resolution system relies on Mitkovs MARS-algorithm (Mitkov et al. 2002), which was introduced in section 2.1.3.

(Poesio 2004) reported an F-measure of 64,2 % for personal pronouns on raw text data of the GNOME corpus (Poesio & Kabadjov 2004). In comparison, the baseline approach (choosing the most recent antecedent) achieved an F-measure of 50,5 % on the same data.

[1] A pleonastic pronoun is non-referential. For example the *it* in "it is raining"

2.1.5 JavaRAP

The JavaRAP algorithm is a anaphora resolution system by (Qiu et al. 2004). It identifies third person pronouns (nominative, accusative or possessive) and lexical anaphors, such as ”himself” or ”myself”.

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 best 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

As already stated, coreference resolution aims for linking all noun phrases referring to the same entity in the real world in a document. 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. Another way of storing coreferences is to define a unique ID for each real-life entity. All occurences in the text will be assigned to their belonging IDs.

An often quoted coreference resolution system using machine learning was proposed by Soon et al. (2001). In this case decision trees was chosen as a classifier. A natural language processing pipeline was used for the identification of markables. The pipeline identified amongst others part-of-speech tags, noun phrases, named entities, and semantic classes. A high value was placed on designing generic features to make them domain-independent. In total, a set of 12 different features was used. 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). Deriving gender information of a noun requires information of their semantic classes. Soon et al. (2001) worked with the simplified assumption 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 was 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. (2001). 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.

The researchers reported an F-measure of 62,6 % on the MUC-6 data and comparable results on the MUC-7 data. A comparison with official MUC-scores indicated, that their system performed at the upper bound of the considered systems. Those values and the used feature set are often referred as baseline for further systems (Versley et al. 2008).

(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 F-measure compared to the initial approach by (Soon et al. 2001).

2.2.2 BART

In 2008, Versley et al. introduced a coreference resolution system for raw text data which extended the previously described approach by Soon et al. (2001). The ambition for BART was to keep it as modular as possible so that it could be applied to many different subtasks of coreference resolution. BART consists of a preprocessing pipeline for parsing, part-of-speech tagging, and further basic information and a mention factory for mainly gender and number identification. Additionaly, a feature extraction module and therefore a matching decoder and encoder is included. The decoder generates the training data while the encoder prepares the testing data. Similar to (Soon et al. 2001) the feature labels are binarized which means that an anaphora either contains the correct or wrong antecedent. Accordingly, the feature labels are either true or false.
A subsequent approach on multiple languages with BART (Broscheit, Poesio, et al. 2010) used a feature set of seven features for all classification types, including a gender agreement, number agreement, string match, and distance feature. The procedure of gaining gender and number information was adopted by Soon et al. (2001).

An F-measure of approximately 55,6 % on Bnews articles of the ACE-2 corpora was reported with the usage of the basic feature set (Versley et al. 2008). With additional language-dependent features, BART was successfully transferred to german (Broscheit, Ponzetto, et al. 2010), polish (Kopec & Ogrodniczuk 2012), and italian (Poesio et al. 2010).

(Reiter et al. 2011) indicated that a great weakness of BART is the implementation of gender information as in their evaluation even noun phrases with explicit gender information were linked incorrectly.

2.2.3 Pronoun Resolution in Spoken Dialogue

As already mentioned, machine learning approaches are less domain-specific than rule-based systems. For that reason (Strube & Müller 2003) presented an corpus-based approach for pronoun resolution in spoken language. Still, several extensions and adaptions had to be done as spoken dialogue differs from written texts gravely. Firstly, the number of pleonastic pronouns in spoken dialogue is substantially increased. Secondly, a not ignorable amount of anaphors in spoken dialogue dont have a clearly defined antecedent so that even humans cant determine them. (Eckert & Strube 2000) called them vague anaphors and figured out that 13.2 % of all anaphors in their examined corpus fall in that category.
A corpus of twenty switchboard dialogues was used. In order to generate training data, a list of all potential anaphors was created. Potential anaphors are all non-definite noun phrases except for first and second person pronouns. Each element in the remaining list forms a pair with every preceding noun phrase that does not disagree in gender, number, or person. If the instances corefer they were labelled P, else N. For all anaphoras without explicit noun phrase antecedents other phrases (for instance verb phrases) in the current last two sentences were used to form pairs.
The feature set with a total of 25 features included noun-phrase features, coreference-level features and spoken dialogue features. Noun-phrase features rely on further preprocessing such as gender, number, or the grammatical function of the anaphora or the antecedent. Coreference-level features could be described as low-level preprocessing features. Those features mainly describe the distance between the antecedent and the anaphora, for instance in words or sentences. The features especially for spoken dialogue contain for instance information on how many noun phrases are located between anaphora and antecedent. A decision tree classifier with 20-fold crossvalidation was applied. (Strube & Müller 2003) reported an F-measure of 47.42 % for the full classifier, including all pronouns and all features.

2.2.4 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

dsd AoneBennettComparisonOfMLMD

2.4 Hybrid Approach

2.5 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).

Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., ... Zanoli, R. (2010). Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 104–107).

Broscheit, S., Ponzetto, S. P., Versley, Y., & Poesio, M. (2010). Extending bart to provide a coreference resolution system for german. In *Lrec.*

Chinchor, N. A., & Sundheim, B. (1995). Message understanding conference (muc) tests of discourse processing. In *Proc. aaai spring symposium on empirical methods in discourse interpretation and generation* (pp. 21–26).

Eckert, M., & Strube, M. (2000). Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics*, *17*(1), 51–89.

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.

Ingria, R. J., & Stallard, D. (1989). A computational mechanism for pronominal reference. In *Proceedings of the 27th annual meeting on association for computational linguistics* (pp. 262–271).

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).

Kopec, M., & Ogrodniczuk, M. (2012). Creating a coreference resolution system for polish. In *Lrec* (pp. 192–195).

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).

Mitkov, R., Belguith, L. H., & Stys, M. (1998). Multilingual robust anaphora resolution. In *Emnlp* (pp. 7–16).

Mitkov, R., Evans, R., & Orasan, C. (2002). A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *International conference on intelligent text processing and computational linguistics* (pp. 168–186).

Mitkov, R., & Stys, M. (2000). Robust reference resolution with limited knowledge: high precision genre-specific approach for english and polish. *Amsterdam studies in the theory and history of linguistic science series 4*, 143–154.

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).

Poesio, M. (2004). The mate/gnome annotation scheme for anaphora deixis, revisited. In *Proc. of sigdial.*

Poesio, M., & Kabadjov, M. A. (2004). A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Lrec.*

Poesio, M., Uryupina, O., & Versley, Y. (2010). Creating a coreference resolution system for italian. In *Lrec.*

Qiu, L., Kan, M.-Y., & Chua, T.-S. (2004). A public reference implementation of the rap anaphora resolution algorithm. *arXiv preprint cs/0406031*.

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.*

Reiter, N., Hellwig, O., Frank, A., Gossmann, I., Larios, B., Rodrigues, J., & Zeller, B. (2011). Adapting NLP Tools and Frame-Semantic Resources for the Semantic Analysis of Ritual Descriptions. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language technology for cultural heritage* (pp. 171–193). Springer.

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.

Strube, M., & Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 1* (pp. 168–175).

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., ... Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Demo session* (pp. 9–12).