

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

1	Introduction	1
1.1	Background	1
1.2	Motivation	2
2	Related Work	4
2.1	Rule-Based Techniques	4
2.1.1	The Naive Hobbs algorithm	4
2.1.2	CogNIAC	5
2.1.3	Anaphora Resolution with Limited Knowledge	5
2.2	Machine Learning-Based Techniques	7
2.2.1	Anaphoras in Coreference Resolution	8
2.2.2	BART	9
2.2.3	Cluster-Based Coreference Resolution	9
2.2.4	Pronoun Resolution in Spoken Dialogue	10
2.2.5	Corpus- and Web-Mined Gender Information	11
2.3	A Comparison of Both Strategies	14
2.3.1	A Manually Designed Resolver (MDR)	14
2.3.2	A Machine Learning-Based Resolver (MLR)	15
2.3.3	Evaluation	15
3	Data	17
3.1	WikiCoref	17
3.2	Gender Corpus	18
4	Methodology	19
4.1	Preprocessing	19
4.2	Feature Set	22
4.2.1	Pronoun Features	22
4.2.2	Antecedent Features	24
4.2.3	Pronoun-Antecedent Features	24
4.2.4	Gender Features	24
4.3	Generating Training Instances	24
4.4	Baseline Approach	24
4.5	SVM Classifier	24
5	Evaluation	25

5.1	Learning curves	25
5.2	Feature Contribution	25
5.3	Error Analysis	25
6	Conclusion	26
6.1	Summary	26
6.2	Outlook	26
	List of Figures	29
	List of Tables	30
	References	31

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 needs 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 considering only those words which get their meaning 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, amongst others 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 for 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 be simplified 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 a 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 anaphoras. The two main purposes are to determine the

impact of gender probability and to compare it to gender information treated as a 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.

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. Since anaphora resolution is a subtask of coreference resolution, coreference resolution systems will be considered as well.

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. In the last section, a comparing evaluation of both techniques will be presented.

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 probable 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 inferior recall (60 %) on their training data consisting of 198 pronouns.

As can be seen the order of rules leads 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.

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 identification module were applied. The algorithm was informally described by Mitkov in three steps:

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

Table 2.1: CogNIAC core rules

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases only to the left of the anaphora
2. Select from the noun phrases identified only those which agree in gender and number with the pronominal anaphora 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). An improved version called MARS had some smaller and greater changes: First of all, three new antecedent indicators and a module for identification of pleonastic pronouns¹ and non-nominal pronominal anaphoras were added. Additionally, the implementation of some previous features was changed as other preprocessing tools were used.

A modular anaphora resolution tool called GuiTAR relying on Mitkovs MARS-algorithm (Mitkov et al. 2002) was developed by 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. (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.

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 fewer 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).

¹A pleonastic pronoun is non-referential. For example, the *it* in “it is raining”

Additionally, machine learning algorithms are usually more time-consuming due to the learning process.

2.2.1 Anaphoras 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. Another way of storing coreferences is to define a unique ID for each real-life entity. All occurrences 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, a decision tree classifier was chosen. A natural language processing pipeline was used for the identification of markables. The pipeline identified among other annotations 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 a 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. Additionally, 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 (Kopeck & Ogronniczuk 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 Cluster-Based Coreference Resolution

The previously described machine learning approaches generate negative and positive training examples as pronoun-antecedent pairs. Rahman & Ng (2009) pointed out several disadvantages of pairwise comparisons: First of all, each possible antecedents is considered on its own which makes a comparison between candidates impossible. For instance, if the first preceding candidate was accepted as the antecedent because it passed a defined threshold, no further candidate will be analysed even if it passed the threshold with a much higher value. Secondly, several contextual information might be missing as only the pronoun-antecedent pair is examined. Those contextual clues could inter alia give information on gender or number. In order to solve those disadvantages Rahman &

Ng (2009) presented a cluster-ranking coreference module. A cluster ranker is trained to determine to which previous coreference cluster a coreference should be resolved. In contrast to rule-based approaches (Mitkov 1998) no manual constraints restrict possible candidates. Instead, restrictions are learned through features automatically. Rahman & Ng (2009) implemented three different classifiers baselines which represent previous learning-based approaches to coreference resolution. For instance, a mention-pair coreference model was implemented. This classifier learns from coreferent-anaphora-pairs which either contain the correct or wrong antecedent (therefore this approach is similar to Soon et al. (2001)). The implemented cluster ranker!!!!!!!!!!!!

When evaluated on the ACE 2005 coreference data sets, cluster rankers outperform three competing models — mention-pair, entitymention, and mention-ranking models — by a large margin.

2.2.4 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 a corpus-based approach for pronoun resolution in spoken language. Still, several extensions and adaptations 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 anaphoras in spoken dialogue don't have a clearly defined antecedent so that even humans can't determine them. (Eckert & Strube 2000) called them vague anaphoras and figured out that 13.2 % of all anaphoras in their examined corpus fall into that category.

A corpus of twenty switchboard dialogues was used. In order to generate training data, a list of all potential anaphoras was created. Potential anaphoras 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 labeled P, else N. For all anaphoras without explicit noun phrase antecedents, other phrases (for instance verb phrases) in the 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 cross-validation was applied. (Strube & Müller 2003) reported an F-measure of 47.42 % for the full classifier, including all pronouns and all features.

2.2.5 Corpus- and Web-Mined Gender Information

Bergsma (2005) presented a machine learning approach to anaphora resolution, which treats gender information not as a hard constraint, but as a probability distribution of possible outcomes. A majority of previous approaches assigned either a specific gender and number (e.g. masculine, feminine, neutral, or plural) or, in case of uncertainty, no gender at all (Soon et al. 2001; Broscheit, Poesio, et al. 2010). Another motivation was that Kennedy & Boguraev (1996) reported to attribute 35 % of their resolution errors to gender mismatch. Only third-person pronouns were considered.

The gender information was derived of two sources: a text corpus and the web. For the former, all occurrences of nouns and pronouns in lexico-syntactic patterns are counted. Five different patterns for reflexives, possessives, nominatives, predicates, and designators were used (Table 2.2). A reflexive masculine occurrence would be for instance “John likes himself”. In this case, a counter for “John” with masculine gender and reflexive pronoun will be increased. This procedure was repeated for all other patterns and remaining genders and numbers (masculine, feminine, neutral, and plural). Bergsma (2005) applied lots of textual data in order to offset parser errors and other noise sources. The whole data set included the AQUAINT corpus (Graff 2002) as well as the Reuters corpus (Rose et al. 2002). In total, a data set of approximately six gigabytes of text was used.

Gender Corpus Indicators	Contained Elements	Pattern
1) Reflexive	himself, herself, itself, and themselves	<i>noun + verb + reflexive</i>
2) Possessive	his, her, its, and their	<i>noun + verb + possessive + noun</i>
3) Nominative	he, she, it, and they	<i>noun + verb + nominative + verb</i>
4) Predicate	he, she, it, and they	<i>pronoun + is/are [a] + noun</i>
5) Designator	Mr. and Mrs.	<i>designator + noun</i>

Table 2.2: Gender Corpus Patterns

Since a text corpus, no matter how big it is, can’t contain all possible words and word combinations, the web was used as a second information source. The Google API was used to count the web pages that appear if a noun, the Google wildcard operator (“*”), and the gender indicator were searched. For instance, if the gender of “John” should be determined, a Google request will be sent with all gender indicating elements of Table 2.2 (John * himself, John * herself, John * itself, etc.). In the following step, the probabilities for each gender will be determined through the five corpus sources and the five web sources. The naive approach would be that the probability of the indicator to be masculine is the percentage of all cases in that the word occurs with its masculine indicator. For instance, in Table 2.3 the cumulated frequency of “Alex” occurring with

“himself” is 60. In total, “Alex” was found 100 times with a reflexive pronoun. As a consequence, the probability for “Alex” to be masculine would be estimated at 60 % from reflexive indicators. This approach leads to three major problems. First of all, zero-probabilities would indicate that there is no possibility for noun to belong to that gender. This might be true - some words might never be part of a certain gender. On the other hand, however, it might just be a rare event and an occurrence would be found with a larger or different text corpus. Secondly, adding a further count could change the likelihood enormous for small frequencies. This leads to the third problem: a measure is needed to determine the certainty of a likelihood. A 70 % probability of a word to be masculine is more meaningful if 1000 cases are considered rather than 10.

In order to solve those problems, Bergsma (2005) treated the counts as a Beta distribution in a Bayesian approach. More precisely, two parameters named α and β are considered. For each gender, α determines the count of the considered event plus one (in order to avoid zero-probabilities) while β represents the count of all not considered events plus one. The α and β values of the previous “Alex” example with reflexive indicators for masculine gender would be $\alpha = 61$ and $\beta = 41$. The mean value of it is computed as:

$$\mu = \frac{\alpha}{\alpha + \beta}$$

A complete distribution is presented in Table 2.3. Note that, unlike the naive approach, these values can only be partially compared to one another, as each of them represents a single distribution. Furthermore, the percentages don’t even sum up to 100 %.

Gender/Number	Occurrences	Naive Approach	Bayesian Approach
1) Masculine	60	$\frac{60}{100} = 60 \%$	$\frac{61}{102} \approx 59.8 \%$
2) Feminine	30	$\frac{30}{100} = 30 \%$	$\frac{31}{102} \approx 30.4 \%$
3) Neutral	10	$\frac{10}{100} = 10 \%$	$\frac{11}{102} \approx 10.8 \%$
4) Plural	0	$\frac{0}{100} = 0 \%$	$\frac{1}{102} \approx 0.1 \%$

Table 2.3: Gender Frequencies Example

Bergsma (2005) expressed the certainty through the variance of Beta distributions:

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

In case of little or no counts at all, the variance will be approximately $1/12$. The classifier should automatically learn that distributions with that variance won't be meaningful.

In order to prove the accuracy of their gender classification, Bergsma (2005) built several SVM-Classifiers. Overall, a set of 20 features was used: Each of the five gender indicators (reflexive, possessive, etc.) has its mean and its variance as features (in this case, the standard deviation was used which is the square root of the variance). Each of the gender indicators was implemented corpus-based as well as web-based. All gender features led to an F-measure of 92 %. Separate classifiers for either web-based or corpus-based information yielded to an F-measure of 85.4 % for the corpus-based and 90.4 % for the web-based approach.

Various pronoun resolution classifiers were built in order to determine the influence of several aspects. In general, each classifier searches, beginning by the certain anaphora, the text backwards until a matching antecedent is found. The matching criteria vary depending upon the complexity of the classifier. The search backwards of the more complex classifiers was limited so that only the current and the previous sentence was considered, because a corpus observation showed that more than 97 % of all antecedents could be found in that range. If no accepted antecedent was found a threshold was reduced so that antecedents with lower likelihood might be accepted. This procedure was repeated until the first candidate exceed the threshold.

The first baseline approach was to always select the most recent noun phrase as antecedent. An accuracy of 26.0 % was reported.

A first improvement consisted of the use of only explicit gender indicators such as "Mr." and "Mrs." to determine the gender. The first previous antecedent that does not mismatch will be chosen. The accuracy was improved up to 30.8%. In a third baseline approach, the previously mentioned gender SVM-classifiers were used to detect a gender match or mismatch. Underlining the importance of gender and number agreement, the accuracy rose up to 59.4 %.

The first machine learning approach included a feature set of 39 features, whereby most of the features were binarized. The features can be separated into three categories. First of all, there are pronoun-related features that determine the gender and number of the pronoun. Secondly, antecedent-related features which provide for instance information on the grammatical relation of the noun phrase or whether it is a person or an organization. The third group of features describes the relation of pronoun and antecedent and contains features that rely on linguistic principles (such as if binding principles are satisfied) as well as features that only require basic preprocessing steps (sentence and word distance, for instance). In order to apply those, the texts were tokenized, parsed, and noun phrases were linked. The training instance creation procedure was adopted by Soon et al. (2001) and was previously described in Section 2.2.1. In total, 1251 positive and 2909 negative training instances were created.

The classifier reached a performance score of 62.3 % which is above all previous ap-

proaches. The additional use of corpus and web frequency features and three other gender affecting features led to a performance score of 73.3 %.

2.3 A Comparison of Both Strategies

Aone & Bennett (1995) did a comparison of a previously build manually designed resolver (MDR) (Aone & McKee 1993) and their in 1995 introduced machine learning-based resolver (MLR).

This section will briefly explain both implementations in order draw an appropriate conclusion of the comparison.

2.3.1 A Manually Designed Resolver (MDR)

The manually designed resolver was build to be language-independent, extensible, robust, and tunable for specific domains. The used information was derived through three different knowledge bases: the *Discourse Knowledge Source*, the *Discourse Phenomenon*, and the *Discourse Domain*.

The former contains antecedent generators to determine all possible antecedents, a system to filter out unwanted antecedent candidates, and an orderer to rank the candidates from highest to lowest likelihood. All of these components rely on specific rules and functions. For instance, the filter removes candidates of mismatching gender. Even though some the rules are only applied on specific languages, (Aone & McKee 1993) reported that most of them are language-independent.

The *Discourse Phenomenon* contains all possible part-of-speech categories in which the anaphora could occur in a hierarchical order. For instance, “third-person” pronoun is a subclass of “pronoun”. Each class includes its definition, two resolution strategies (a second one is needed if the main strategy fails), and specific language information if a category only exists in a certain language.

The third knowledge base is responsible for domain-specific information.

A module called *Discourse Administrator* was used to determine the application domain and in a further step to select and filter the knowledge bases in order to generate the best possible resolution system. Therefore, the information stored in each knowledge base is heavily dependent on the considered language and domain. The general resolution process is as follows: The discourse phenomena are used to determine all anaphoras. In a second step, the discourse knowledge sources are applied in order to generate and filter all possible candidates. If only one remains, it will be chosen as antecedent. Otherwise, one or more orderers are applied and the best candidate will be chosen by order. If no candidate was found at all, the second strategy specified in the discourse phenomenon will be applied.

2.3.2 A Machine Learning-Based Resolver (MLR)

The machine learning-based resolver presented by Aone & Bennett (1995) used pairwise training examples containing information on the anaphora and its possible antecedent. A whole set of 66 features was used. Aone & Bennett (1995) divided most of them into one of four subcategories, namely lexical, syntactic, semantic, and positional. The feature selection inspired by the manually designed resolver (Aone & McKee 1993), but were generalized and changed in order to be domain- and language independent.

In total, six different classifiers depending on three parameters were trained. The first parameter was called anaphoric chain. If its value is true, a correct antecedent is detected if the candidate is part of the same anaphoric chains which means that both refer to the same real-world entity. Otherwise, just the preceding same-world entity will be accepted as correct antecedent. This parameter also affects the training instance generation. In case of anaphoric chains, all co-referring phrases will form positive training instances with its anaphora. In the other case, just the preceding co-referring phrase will be used for positive instances. In both cases, the remaining phrases will form negative training instances with the anaphora. The second parameter determines whether the decision tree will use further information of the anaphoric type (for instance whether the real-world entity of the anaphora is a proper name). A third parameter determines the pruning-factor of its decision tree. A high pruning-factor indicates a higher generalization while decision trees with a lower factor tend to overfit.

2.3.3 Evaluation

The comparison was evaluated on japanese newspaper articles. In total, 1271 anaphoras were used. As it can be seen in Table 2.4, all machine learning approaches using anaphoric chains outperformed the manual approach independent of their pruning-factor, while the approach without the usage of anaphoric chains performed slightly worse. The different pruning factors seemed to have a rather low impact on the performance.

As the manually designed resolver also detects only the preceding same-world entity, it would be most reasonable to compare it to the MLR-6. Even though the manual approach performed better, no language specific information or relevance of features need to be determined as the algorithm learned it autonomously (Aone & Bennett 1995). Aone & Bennett (1995) interpreted the results as optimistic for machine learning techniques.

Algorithm	Anaphoric Chains	Anaphoric Type	Confidence	F-measure
MLR-1	yes	no	100 %	76.27
MLR-2	yes	no	75 %	77.30
MLR-3	yes	no	50 %	76.43
MLR-4	yes	no	25 %	77.28
MLR-5	yes	yes	75 %	74.54
MLR-6	no	no	75 %	67.03

MDR	69.57
-----	-------

Table 2.4: Aone & Bennett Evaluation

Chapter 3

Data

Most machine learning approaches require annotated corpora in order to make the learning process possible. In this case, the training corpora must contain information on the correct antecedent for each anaphora. Since this work is designed to learn gender information through frequencies, a second information source is needed. This chapter will briefly describe both information sources.

3.1 WikiCoref

Ghaddar & Langlais (2016) presented with WikiCoref a coreference-annotated corpus of english Wikipedia articles. Wikipedia differs from most web corpora as it is highly structured. The Wikipedia guidelines¹ contain various restrictions on grammar and vocabulary and also define the structure of articles in terms of sections and paragraphs. In contrast, most other web-mined sources are heavily unstructured and could contain colloquial language as well as ungrammatical text.

An excerpt of 30 articles was used to build the corpus. (Ghaddar & Langlais 2016) figured out that more than 35 % of all Wikipedia articles contain less than 100 words and only 11 % more than 1000 words. Articles with few word counts (less than 200 words) were not considered as they don't contain enough information for meaningful coreference resolution. Hence articles can't be chosen completely random, a uniform distribution of categorized article sizes leading from less than 1000 to more than 5000 words was strived. Additionally, articles with too many out links were not considered. In order to keep the corpus domain-independent, articles of different topics were selected.

To detect entities, a combination of a coreference resolution system, an entity detection module, and anchored links in the article was used. The coreference chains detected by the module were manually corrected and missing ones were added. All coreferring entities were linked through a joint id representing the real-world entity. For instance:

¹https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[Aberfoyle]₁ is [a village in the region of Stirling, Scotland, northwest of [Glasgow]₂]₁.

[The town]₁ is situated on [the River Forth]₃ at the base of [Craigmore]₄ (420 metres high).

In total, the corpus contains 59652 tokens² in 2229 sentences with an average of 2000 tokens per document. For the inter-annotator agreement an MUC F1 score (Vilain et al. 1995) of 83,3 % was reported.

3.2 Gender Corpus

An automatic approach to learning gender information through corpus- and web-based frequencies was introduced by Bergsma (2005) and explained in Section 2.2.5 of this work. Bergsma & Lin (2006) pointed out two disadvantages of the previous approach. First of all, sending Google requests for each possible antecedent on large corpora is time-consuming and therefore not cost-efficient. Secondly, the corpus- and web-based implementations are not symmetric as some occurrences can only be found with the web-based approach. For instance, the corpus-based approach merely accepts a verb between a noun and a reflexive pronoun while the Google wildcard operator ("*") is not limited to any grammatical category. Therefore, a new corpus mined frequency distribution of gender and number information was mined using a corpus of approximately 85 GB. Overall, an accuracy of 90,3 % on gender determination was reported. Bergsma & Lin (2006) made the mined gender and number frequencies openly accessible for the NLP community.³

²"A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing." (<http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html>)

³Available for download at <http://www.clsp.jhu.edu/~sbergsma/Gender/Data/>

Chapter 4

Methodology

This chapter will present the whole procedure of the implemented anaphora resolution system with all of its stages. The complete system can also be examined online.¹

4.1 Preprocessing

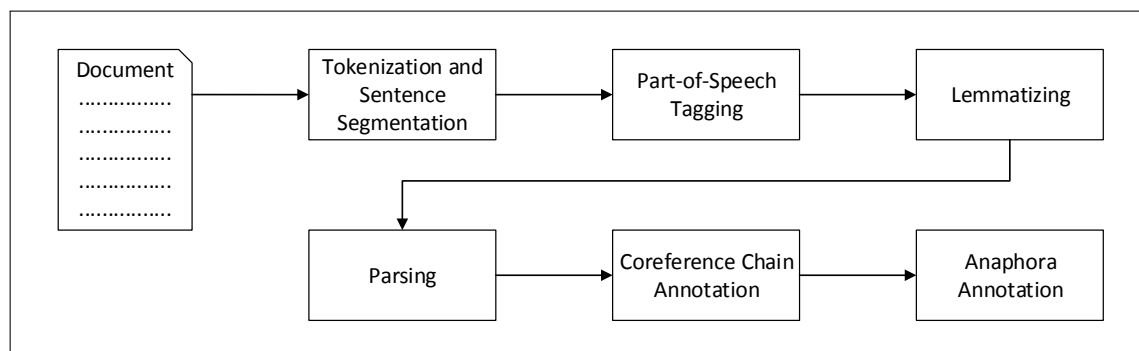


Figure 4.1: Natural language preprocessing pipeline

First of all, the required information of the training corpus need to be extracted. The natural language preprocessing pipeline shown in Table 4.1 was used. The WikiCoref annotation scheme already includes information on tokens, sentences, and coreferential chains and could easily be extracted.

Still, several other information is missing in order to apply feature values (for instance, information on nominal phrases and part of speech). The *Stanford CoreNLP* toolset (Manning et al. 2014) was used to gain those informations. More precisely, its Part-of-Speech tagger, lemmatizer, named entity recognizer, and parser were applied. Note that the assigned labels for the part-of-speech tagger and the parser are simplified in order to reduce complexity of the following examples. For instance, the implemented

¹The download is available at <https://github.com/HenryvanderVegte/henryvdv.BA>

part-of-speech tagger will differentiate between A Part-of-Speech tagger annotates for each token a word class. Word classes are for instance nouns or verbs. Additionally, the tagger differentiates also on more specific details like number or tense. In total, the tagset contains 52 different tags. The lemmatizer creates for each word its canonical form. A lemma is similar to the word stem, but focuses of the use of the word while most stemmers mostly remove the end of the word to receive its word stem². The named entity recognizer identifies amongst others persons, organizations, and dates. An illustrative example of these elements is shown in Figure 4.2.

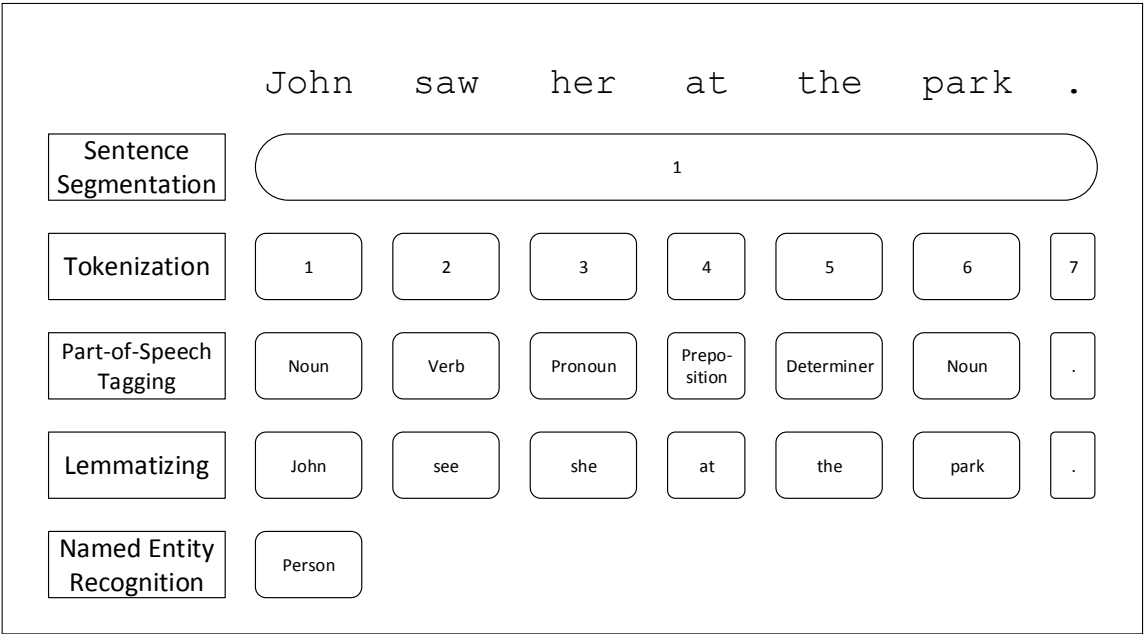


Figure 4.2: Annotation example

The Parser can be subdivided in two different modules named constituent parser and dependency parser.

The constituent parser divides the sentence into sub-phrases in an hierarchical order. The type of the phrase is defined by the central word in it (also called head word) (Jurafsky & Martin 2014). For instance, if the head word is a noun the considered phrase is called noun-phrase. The most common kind of illustration is through a dependency parsed tree as shown in Figure 4.3.

The dependency parser describes the relationship of words among each other. A word that is dependent of another is linked directed to it. Additionally, the relationship between both words is annotated. Most graphic representation visualize the relationship through pointed arrows based on the dependent word. An exemplary dependency parsed model is shown in Figure 4.4. The arrow description will thereby represent the type

²<http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

of the relationship. The abbreviations subj., mod., obj., and det. represent subjects, modifiers, object, and determiners.

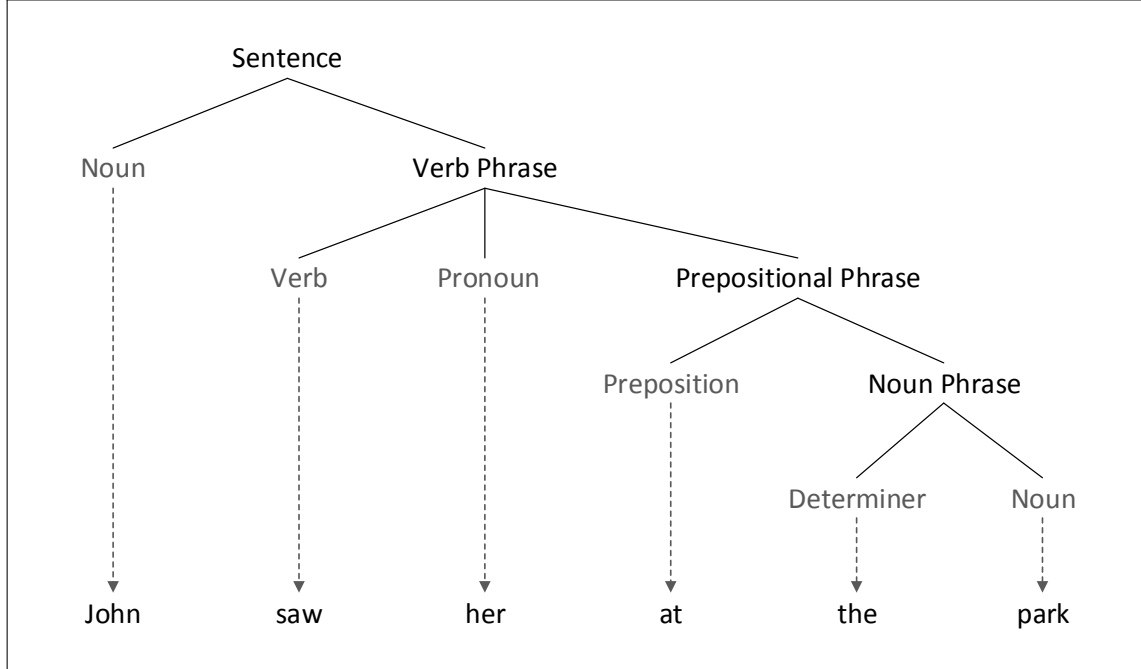


Figure 4.3: Constituency-based parse tree example

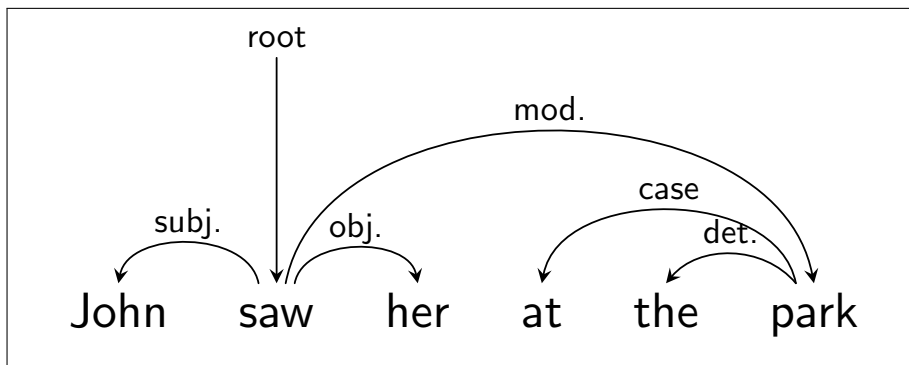


Figure 4.4: Dependency parsed example

Coreferential information in WikiCoref is stored through a joint identification number for each coreferring annotation and can therefore be easily extracted. The third person pronominal anaphoras were extracted of the coreference chain as follows: For each pronoun in the current document, search the preceding phrase of its coreference chain. The pronoun will be tagged as the anaphora and the preceding phrase as the antecedent.

As soon as all relevant information is annotated, a training set needs to be created.

In order to do so, the procedure Soon et al. (2001) was adopted. As already stated in Section 2.2.1, all anaphoras and their respective antecedents form thereby positive feature vectors while all intermediate noun phrases form negative feature vectors with their respective anaphoras. For instance, a sequence of noun phrases A-B-C-D with A coreferring D and D as pronoun is considered. In this case the pair (A,D) forms a positive instance while (C,D) and (B,D) form negative instances.

The idea of selecting that approach is as follows: A human would reject all intermediate noun phrases due to several indicators that exclude the candidate for some reason. In contrast, noun phrases earlier in the text than the antecedent could be legitimate candidates if the antecedent would not exist. In conclusion, a valid rejection of candidates preceding the antecedent is only possible if the antecedent is already detected. However, if the antecedent is already detected no further examination needs to be done. With the algorithm of Soon et al. (2001), 904 positive and 2164 negative were created in total.

The anaphora set consists of 17 reflexives (herself, himself, themselves, itself), 379 nominatives (he, she, they), and 508 possessives (their, its, her, his). An "it" appearing in structures like "it is raining" is considered pleonastic. There are several ways of dealing with that non-referential structures, leading from machine learning-based (Boyd et al. 2005) to rule-based approaches (Lappin & Leass 1994). Several pronoun resolution approaches decided to manually exclude pleonastic pronouns (Kennedy & Boguraev 1996; Bergsma 2005). As the number of occurrences of "it" is not decisive, it will be neglected in this work.

4.2 Feature Set

The whole feature set can be divided in four different categories. There are features that only affect the anaphora, features that only affect the antecedent, features that describe the relationship between both, and features that are related to gender information. A whole list of features is shown in Table 4.1. The features are mostly adopted by Bergsma (2005), however there are several different implementations especially in the gender features.

4.2.1 Pronoun Features

As the list of resolved pronouns is limited, a simple rule for each feature determines the gender values (for instance, if the pronoun is whether "he", "his", or "himself", it is considered as masculine).

Feature Type	Feature	Description
Pronoun referred	Masculine	If pronoun is masculine: 1, else: 0
	Feminine	If pronoun is feminine: 1, else: 0
	Neutral	If pronoun is neutral: 1, else: 0
	Plural	If pronoun is plural: 1, else: 0
Antecedent referred	Antecedent Frequency	number of occurrences / 10.0
	Subject	If antecedent contains subject: 1, else: 0
	Object	If antecedent contains object: 1, else: 0
	Predicate	If antecedent contains predicate: 1, else: 0
	Pronominal	If antecedent contains pronoun: 1, else: 0
	Head-Word Emphasis	If antecedent parent is no noun: 1, else: 0
	Conjunction	If antecedent is not part of conjunction: 1, else: 0
	Prenominal Modifier	If antecedent contains prenominal modifier: 1, else: 0
	Organization	If antecedent contains organization: 1, else: 0
	Person	If antecedent contains person: 1, else: 0
	Time	If antecedent contains time units: 1, else: 0
	Date	If antecedent contains date: 1, else: 0
	Money	If antecedent contains monetary name: 1, else: 0
	Number	If the antecedent contains a number: 1, else: 0
	Definite	If antecedent has definite article: 1, else: 0
	His/Her	If antecedents first word is his or her: 1, else: 0
	He/His	If antecedents first word is he or his: 1, else: 0
Pronoun and antecedent referred	Binding Theory	If binding principles B and C are satisfied: 1, else: 0
	Same Sentence	If both are in the same sentence: 1, else: 0
	Intra-Sentence Diff.	Difference in sentences / 50.0
	In Previous Sentence	If antecedent is in previous sentence: 1, else: 0
	Inter-Sentence Diff.	Difference in tokens / 50.0
	Prepositional Parallel	If both depend on the same preposition: 1, else: 0
	Quotation Situation	If both are in or out quotes: 1, else: 0
	Singular Match	If both are singular: 1, else: 0
	Plural Match	If both are plural: 1, else: 0
Gender referred	Standard Gender Match	If gender is known and matches: 1, else: 0
	Standard Gender Mismatch	If gender is known and mismatches: 1, else: 0
	Pronoun Mismatch	If both are pronouns and mismatch: 1, else: 0
	Masculine Mean	μ of masculine distribution
	Masculine Variance	Variance of masculine distribution
	Feminine Mean	μ of feminine distribution
	Feminine Variance	Variance of feminine distribution
	Neutral Mean	μ of neutral distribution
	Neutral Variance	Variance of neutral distribution
	Plural Mean	μ of plural distribution
	Plural Variance	Variance of plural distribution

Table 4.1: Pronoun Resolution Feature Set

4.2.2 Antecedent Features

sds

4.2.3 Pronoun-Antecedent Features

4.2.4 Gender Features

4.3 Generating Training Instances

4.4 Baseline Approach

4.5 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

4.1	Natural language preprocessing pipeline	19
4.2	Annotation example	20
4.3	Constituency-based parse tree example	21
4.4	Dependency parsed example	21

List of Tables

2.1	CogNIAC core rules	6
2.2	Gender Corpus Patterns	11
2.3	Gender Frequencies Example	12
2.4	Aone & Bennett Evaluation	16
4.1	Pronoun Resolution Feature Set	23

References

- Aone, C., & Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd annual meeting on association for computational linguistics* (pp. 122–129).
- Aone, C., & McKee, D. (1993). A language-independent anaphora resolution system for understanding multilingual texts. In *Proceedings of the 31st annual meeting on association for computational linguistics* (pp. 156–163).
- 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).
- Bergsma, S., & Lin, D. (2006, July). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics* (pp. 33–40). Sydney, Australia: Association for Computational Linguistics.
- Boyd, A., Gegg-Harrison, W., & Byron, D. (2005). Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the acl workshop on feature engineering for machine learning in natural language processing* (pp. 40–47).
- Broscheit, S., Poesio, M., Ponzetto, S. P., Rodriguez, K. J., Romano, L., Uryupina, O., ... Zanolini, 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).
- Ghaddar, A., & Langlais, P. (2016, 05/2016). Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).
- Graff, D. (2002). The acquaint corpus of english news text. *Linguistic Data Consortium, Philadelphia*.
- 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).
- Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing*. Pearson.
- Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: pronominal anaphora resolution 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).
- Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4), 535–561.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for computational linguistics (acl) system demonstrations* (pp. 55–60). Retrieved from <http://www.aclweb.org/anthology/P/P14/P14-5010>
- 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*.
- Rahman, A., & Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 2-volume 2* (pp. 968–977).
- 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.
- Rose, T., Stevenson, M., & Whitehead, M. (2002). The reuters corpus volume 1-from yesterday’s news to tomorrow’s language resources. In *Lrec* (Vol. 2, pp. 827–832).
- 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).
- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on message understanding* (pp. 45–52).