MASTER'S THESIS | LUND UNIVERSITY 2017

Creating a coreference solver for Swedish and German using distant supervision

Alexander Wallin

Department of Computer Science
Faculty of Engineering LTH

ISSN 1650-2884
LU-CS-EX 2017-03

Creating a coreference solver for Swedish and German using distant supervision

(Training coreference solvers using machine annotated data)

Alexander Wallin
alexander@wallindevelopment.se

March 2, 2017

Master's thesis work carried out at
the Department of Computer Science, Lund University.

Supervisor: Pierre Nugues, pierre.nugues@cs.lth.se

Examiner: Jacek Malec, jacek.malec@cs.lth.se

Abstract

Coreference resolution is the identification of phrases that refer to the same entity in a text. Current techniques to solve coreferences use machine-learning algorithms, which require large annotated data sets. Such annotated resources are not available for most languages today. In this report, we describe a method for solving coreference for Swedish and German without annotated texts using distant supervision. We generate a weakly labelled training set using multilingual corpora, where we solve the coreference for English using CoreNLP and transfer it to Swedish and German using word alignment. Additionally, we identify mentions from dependency graphs in both languages using handwritten rules. Finally, we evaluate the end-to-end results using the evaluation framework from the CoNLL 2012 shared task where we obtain an F-measure of 34.98 for Swedish and 13.16 for German.

Keywords: coreference resolution, distance supervision, machine-learning, multilingual, Swedish, German

Acknowledgements

I would first like to thank my supervisor, Pierre Nugues, for broaching the subject of multilingual coreference resolution and his time and effort to see it through. Pierre has served as an inspiration throughout my academic career and his tireless support and counsel factors heavily into my interest in the field of natural language processing.

I am also grateful for my examiner, Jacek Malec, whom I have always held in the highest regard. He was a close second when I considered topics and supervisors for my Master's thesis.

I would also like to thank Marcus Klang, Peter Exner, and Dennis Medved for their advice and counsel.

It is also certainly warranted to thank Christian Lindgren at some point, as he was the catalyst for my interest in NLP.

Finally, I would also like to thank my family; without whom my achievements would be for nought and nothing would ever matter. I dedicate this work to them; family forever first.

Contents

Chapter 1
Introduction

Coreference resolution is the process of determining whether two expressions refer to the same entity and linking them in a body of text. The words and phrases relevant for coreference resolution are generally referred to as *mentions*, a sentence fragment that mentions a referent.

For a human reader coreference resolution is a subconscious aspect of reading; when reading a text such as:

> Alexander works on his Master's thesis. He enjoys the challenge.

it may be assumed that *Alexander* and *He* refer to the same real world entity, namely the thesis' author.

If a machine were to analyze the same text it could possibly deduce that a named person is working on a Master's thesis and that a male person enjoys a challenge, but would not by necessity see a casual link between the two as their relationship is inferred from contextual information – the distance between the two sentences, that *Alexander* and *he* share the same multiplicity and gender.

If the sentence were to read *Alexander works on his Master's thesis. They enjoy the challenge.* a human reader would react with confusion and wonder whether the text was only a fragment of a larger text or if the writer had made an error. This is because *they* refer to a group of people of unknown gender while *he* refer to one person. Since the gender and multiplicity of *he* and *Alexander* is in agreement the two mentions a human reader would find it likely that they refer to the same entity.

If someone were to be interested in extracting information from the text they would be able to infer that a named person (Alexander), who is likely a man, is working (on his thesis) and that a male person enjoys a challenge. By linking *he* with *Alexander*, which is the purpose of coreference resolution, it would be possible to infer that a named male person enjoys the challenge of writing his thesis. A simple graphical representation of the information extracted from the two sentences is shown in Fig 1.1, where the information read from the sentences is represented by the two entities on the left. By identifying that

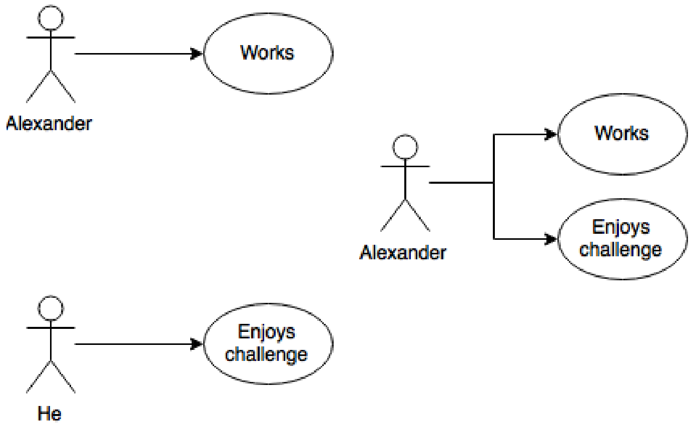

Figure 1.1: Deductions from using coreferrence

the two entities are in fact identical, it would be possible to deduce the entity on the right side of the graphical representation.

These two sentences are as simplistic as they are whimsical, but many texts are neither and the lack of coreference resolution tools hinders qualitative knowledge extraction.

1.1 Aim of the thesis

The goal for this Master's thesis is to create machine annotated corpora with annotated coreference chains in Swedish and German using the multilingual Europarl corpus. The corpora are then used to train coreference resolvers using distantly supervised learning methods. The evaluation of coreference solving can be tricky. The CoNLL 2011 and CoNLL 2012 conferences defined a procedure that is widely accepted in the field. The trained coreference resolvers will be evaluated on hand annotated data using the evaluation procedure defined by the CoNLL 2011 and CoNLL 2012 conferences. The Swedish coreference solver will most likely be one of the first commonly available solvers for the language.

1.2 Background

For a few languages there exists easily available coreference resolvers with good performance, but this is far from the case for every language. Some languages such as the Nordic languages lacks, to the best of our knowledge, tools for solving coreference.

The reason for this is lack of available training data; the current state of the art utilizes machine-learning to train solvers using large hand-annotated corpora with its coreferences resolved. Large hand-annotated corpora with resolved coreferences require tremendous effort to produce, which is too expensive for most languages.

In the case of Swedish the number of commonly available hand corpora with annotated with coreference is limited. To the best of our knowledge there exists only one; SUC-Core. SUC-Core consists of 20,000 words and 2,758 coreferring mentions, which is generally considered small when training a coreference solver.

In comparison, the CoNLL 2012 shared task uses a training set consisting of more than a million words and 155,560 coreferring mentions for the English language alone and this is only a subset of the available resources for this particular language (Pradhan et al., 2012).

Models trained on large corpora do not automatically result in a better model, but one can presume that the two orders of magnitude difference between the size of the English CoNLL 2012 corpora and SUC-Core would yield a more universal model. Pradhan et al. (2012) propose that larger and more consistent corpora and the use of a standardized evaluation scenario would likely improve on the results of coreference resolution where those measures are not implemented. It would therefore be easy to assume that a large and consistent corpora would benefit coreference resolution for the Swedish language.

1.3 Distant Supervision

Distant supervision is a form of supervised learning, though the term is sometimes used interchangeably with *weak supervision* and *self training* depending on the source (Yao et al., 2010).

The primary difference between distant supervision and supervised learning lies in the quality of its training data; supervised learning uses labelled data, whereas in the case of distant supervision the training data is generated. Training data can be generated using various methods, such as simple heuristics or train a model using the output from another model. Distant supervision will often yield models that perform less well than models using other forms of supervised learning (Yao et al., 2010). The strength of the distant supervision method comes from introducing supervised learning methods to problems without labelled data. There is a wide range of methods which constitutes distant supervision, depending on context. The type of distant supervision used for this Master's thesis is annotation projection, where the output from a different coreference resolver is transferred using a bilingual corpora and used as input for training a solver in another language (Martins, 2015; Exner et al., 2015).

1.4 Corpora Creation

The problem with consistently annotating a large corpora is the time and effort required to manually annotate and evaluate large bodies of text. Consider the SUC 2.0 corpora as an example of this notion. In the introduction to the SUC 2.0 manual the authors wrote *never more* as an epitaph for the time and effort spent collecting and annotating the corpus Gustafson-Capková and Hartmann (2006).

Rahman and Ng (2012) use advances in statistical machine translation to align words and sentences from a language to another and transfer annotated data and other entities from one language to another. Their solution was to collect a large corpus of text in Spanish and Italian, translating each sentence using machine translation, applying a coreference solver on the generated text and aligning the sentences using unsupervised machine translation methods.

Hwa et al. (2005) use parallel texts to transfer syntactic annotation using a parallel language corpus. The method described by Rahman and Ng (2012) use machine translation which seems like an unnecessary step if a multilingual corpus were available. We introduce a method were we use a parallel corpus as with Hwa et al. (2005), but with the methods and metrics described by Rahman and Ng (2012). This would allow for the creation of a large and consistently annotated labelled corpus, the addition of systemic errors stemming from machine annotation and incorrectly transferred entities using the word alignment when compared with an equivalently sized hand-annotated corpora notwithstanding. Annotating a large corpora in Swedish by hand would, after all, be unfeasible for this project.

1.5 Selection of Corpora

The Europarl corpus is a large multilingual corpus containing protocols and articles from the EU parliament gathered from 1996 to the present in the various languages spoken in the parliament. The corpus aligns sentences from one language to another but is otherwise unlabeled. The corpus has proven useful for multiple natural language processing tasks in the past (Koehn, 2005) and was therefore an ideal candidate for corpora selection. It consists of 21 language pairs with hundreds of thousand sentences in each language, which is well beyond the minimal requirements for training a general coreference solver. Aligning sentences from one language to another can be done either by hand or by using machine-learning. Europarl uses a machine-learning algorithm based on unsupervised learning to align sentences. Unsupervised learning uses word correlation and patterns to induce sentence alignments from one language to another. This method is not ideal, as human alignment would be preferable, but should be sufficient for our purposes, as incorrectly aligned sentences will likely result in poor word alignments that are likely to be rejected by our method.

1.6 Standardized Evaluation Scenario

The CoNLL 2011 and CoNLL 2012 shared tasks focused on coreference resolution in three specific languages; English, Arabic and Chinese using a standardized evaluation tool and metrics (Pradhan et al., 2011, 2012). We used SUC-Core as a test set and by using the tools and metrics from the CoNLL shared task a standardized evaluation scenario for Swedish can be devised (Nilsson Björkenstam, 2013).

In the case for the German language we used the Tüba-D/Z corpus for evaluation in the same manner as SUC-Core (Henrich and Hinrichs, 2013, 2014).

1.7 Related Work

This Master's thesis utilizes various natural language processing subtasks and is heavily dependent on the work and research of people from a broad range of topics.

Martins (2015) developed a coreference solver for Spanish and Portuguese using distant supervision, where he transferred entity mentions from English to a target language using machine-learning techniques. The article presents a complete end-to-end system for transferring and training a coreference solver using distant supervision. The concepts we reused were primarily the use of the maximum span heuristic and the pruning of documents according to the ratio between correct and incorrect entity alignments. The maximum span heuristic was originally introduced by Yarowsky et al. (2001), but for another purpose than that of Martins (2015).

Stamborg et al. (2012) built a coreference solver using the original algorithms devised by Soon et al. (2001). They obtained high accuracy through an optimized feature set. This Master's thesis reused a subset of their feature set and their use of dependency trees.

The Europarl corpus is based on a large collection of documents in a large range of languages spoken in the European Union. Documents in one language often has documents with similar or identical content in other languages. The research team behind the Europarl corpus has identified matching documents using machine-learning methods specifically developed for Europarl. The corpus also contains matching sentences, which were identified using the Gale and Church algorithm (Koehn, 2005). Even though the corpus identifies two sentences as matching they are sometimes only barely similar, which complicates aligning words in one language to another. By using the precalculated word alignments from the open parallel corpus OPUS many of the challenges stemming from improper word alignment was mitigated (Tiedemann, 2012).

Rösiger and Kuhn (2016) created a data-driven coreference model for the German language based on the German hand-annotated Tüba-D/Z corpora. They adapted the latent tree coreference solver IMS-HotCoref by Björkelund and Kuhn (2014) to work on the German language. As IMS-HotCoref was originally devised to solve coreference for English, Arabic, and Chinese, their paper was very helpful when adapting our model to German.

1.8 Terminology

Ideally every concept and phrase in this Master's thesis would be sufficiently explained in the other sections, but this would be both infeasible and impractical. Most of these concepts and phrases are instead presented as a list in this section with a short explanation.

Anaphora and Antecedent *Anaphora* and *antecedent* are used throughout this report to describe the relationship between two mentions. There are various definitions for these terms, which would make interpreting them in this report somewhat obtuse, and is therefore explained in greater detail.

Anaphora is a mention that refers back to a previous mention that is called the antecedent. In cases where the relationship between two mentions is not known, for instance before solving the document's coreference, this report uses the terms to describe their relative order in the document - antecedent is first, followed by the anaphora.

Word or Phrase	**Explanation**
NLP	Short for Natural Language Processing
Coreference Solver	The software that solves for coreference
Coreference Resolution	The results from the solver
Corpus	NLP term for a data set that (usually) contains text, can also be annotated.
Dependency Graph	A representation of a sentence in the form of a directed graph based on the dependency relation.
UTF-8	A common character encoding format for text.
CoNLL	A large top tier conference in the field of natural language processing

Chapter 2
Approach

The goal for this Master's thesis is the creation of a coreference solver for Swedish and German without using labelled training data. the Swedish language lacks corpora of sufficient size to train a general coreference solver, whereas the German language has a large labelled corpus in the form of Tüba D/Z which has been successfully used to train coreference solvers in the past. Tüba D/Z is discussed in more detail in Sect. 4.4. Our approach was primarily developed with the Swedish language in mind, as the German language already has sufficient resources with existing corpora to train solvers. The German language was selected as to allow us to evaluate our method's general applicability.

There are various ways to train models without using labelled data. For this thesis we use distant supervision, which is a form of supervised learning trained on weakly labelled data.

What constitutes weakly labelled data is somewhat open for interpretation, but it generally means data labelled without human interaction.

The weakly labelled data was generated by using a corpus consisting of sentence-aligned text with sentence mapping from English to Swedish and English to German. The English text was annotated using a coreference solver for English and the coreference chains were then transferred to the target language by word alignment. The transferred coreference chains were then used to train coreference solvers for the target language. This section discusses the generation of the weakly labelled training data, while Sect. 5 discusses the implementation details for the coreference solver.

2.1 Europarl Corpus

Europarl is a large sentence-aligned unannotated corpus based on the proceedings of the European Union from 1996 to the present day (Koehn, 2005). The corpus consists of both text documents and web data in the XML format, but only text documents were used for this study. Koehn evaluated the Europarl corpus using the BLEU (*bilingual evaluation* *understudy*) metric, which evaluates the quality of machine translated text (Papineni et al., 2002)(Koehn, 2005). High BLEU scores are preferable as it often results in better word alignments (Yarowsky et al., 2001).

The BLEU metric has a scale from 0 to 100 where 0 means no alignment and 100 means perfect alignment. The values for Europarl ranged from 10.3 to 40.2, with the English-to-Swedish at 24.8 and English-to-German at 17.6.

Additionally, Ahrenberg (2010) notes that the English-Swedish alignment of Europarl contains a high share of structurally complex relations which makes word alignment more difficult (Ahrenberg, 2010).

The documents in the Europarl corpus are downloaded in files encoded in UTF-8 with sentences ending at every linebreak. In addition to these language specific documents each language pair has alignment files which maps which sentences maps to each other in the different languages. As there is no information which can be transferred from a sentence which does not map to another sentence these unaligned sentences were removed from every document.

2.2 Machine Annotation

We used language-dependent processing pipelines to annotate our texts:

- The English text was annotated with parts of speech, dependency grammar, and coreference using Stanford's CoreNLP (Manning et al., 2014).
- The German text was annotated using Mate Tools developed by researchers from both IMS Stuttgart and LTH (Björkelund et al., 2010).
- The Swedish text was annotated with part of speech using Stockholm University's Stagger and annotated with dependency grammar using Växsjö and Uppsala University's MaltParser (Östling, 2013; Nivre et al., 2007).

Special attention on the sentence boundaries for all parsers was required as to not differ from the boundaries defined by Europarl, as these boundaries define the boundaries for the word alignment. The recommended method described in Stanford's CoreNLP was to turn off the sentence boundary check and automatically end each sentence with the newline character. These steps were dully executed for all relevant languages.

The following subsections contain detailed descriptions of the tools used for machine annotating the sentences.

2.2.1 CoreNLP

Stanford's CoreNLP is available under open source licenses while also being a high performing end-to-end annotator for English and other languages which includes models for coreference. This high performing annotator was used by a team from Stanford for the CoNLL 2011 shared task and obtained the best results for English coreference (Pradhan et al., 2011). The latent tree approach performed better at the CoNLL 2012 shared task, but we still decided to use CoreNLP as its readily available with pretrained models (Pradhan et al., 2012). In addition to annotating English, it has models for German. It lacks models for solving coreference in German, but it is able to annotate German parts of speech and dependency grammar.

2.2.2 Stagger and MaltParser

Stagger, developed by Östling (2013) at Stockholm University, is an open source part of speech tagger based on Collins (2002) Averaged Perceptron. The tagger has a commonly available model for the Swedish language based on the SUC corpus (Gustafson-Capková and Hartmann, 2006).

MaltParser, developed by Joakim Nievre et al. provides robust dependency parsing for various languages. The Swedish model is based on the Talbanken section of the Swedish Treebank, which is a small but freely available subset of SUC (Gustafson-Capková and Hartmann, 2006).

Stagger and MaltParser are some of the few annotators for the Swedish language with good performance characteristics and readily available models.

2.2.3 Malt Tools

Malt Tools is a toolkit for end-to-end annotation developed jointly by researchers from both IMS Stuttgart and LTH (Björkelund et al., 2010). The toolkit has pretrained models for various languages, but only German was used for this thesis. The German model is trained on the Tiger corpus (Brants et al., 2004).

2.3 Word Alignment

The premise for this thesis is the proper transfer of entity mentions from one language to another. As such, correctly aligning the English words with the Swedish and German words is of key importance. This section describes the process and the algorithms used throughout the project, from start to finish.

In terms of statistical machine translation the correct term for English in this thesis is the *source language*, as information is transferred from English to either Swedish or English. Analogous, the Swedish and German languages are called the *target languages*, as they are targets of the source language.

2.3.1 IBM Alignment Models

IBM is probably best known today for Watson (Ferrucci et al., 2010), but they have actually written some of the most widely used algorithms for statistical machine translation, namely the IBM Alignment Models (Brown et al., 1990, 1993).

The IBM algorithms is a set of five incrementally developed algorithms for aligning words from one language to another in multilingual corpora. The resulting links are directed graphs where every target word is linked to at most a single source word.

The algorithms were reimplemented for the open source Giza system (Al-onaizan et al., 1999), which in turn saw improvements into Giza++, an open source implementation of the original Giza code and the first four IBM models (Och and Ney, 2000).

2.3.2 Word Alignment using Heuristics

The sentences for each document were first tokenized and lemmatized. The words in each sentence were then aligned using the fourth IBM alignment model implemented in Giza++.

The quality of the alignment varied heavily, especially for semi-unique words such as proper nouns and numerals, which given their infrequency were mapped incoherently.

Some common errors can be viewed in Figs. 2.1 and 2.2. In Fig. 2.1, *santos* is not mapped to *Santos*, while the word *part-session* is incorrectly mapped to *frågestunden under sammanträdesperiod* instead of only *sammanträdesperioden* in Fig. 2.2.

We proposed a unique solution to this problem, where pronouns are first mapped according to string equivalence and only afterwards according to the results from Giza++.

Additionally, when the number of proper nouns in the source and target sentences differed the sentences were ignored in subsequent data processing. After sampling a subset of the aligned documents we deemed this method insufficient as the number of aligned mentions was low. We decided to use the word alignments from the OPUS corpus instead.

An example of the results for an idealized sentence is shown in Fig. 2.3.

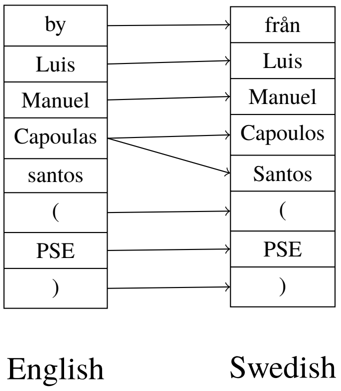

Figure 2.1: Example of incorrect projection

2.3.3 OPUS

Word alignment according to the heuristic approach described in Sect. 2.3.2 was, despite an improvement on the initial word alignment results, insufficient for the task at hand. We decided to utilize the Europarl corpus contained in the open parallel corpus OPUS, as this corpus contains precalculated word alignments. The word alignments in OPUS uses the phrase based grow-diag-final-and heuristic, which gave more well-behaving results than the initial endeavours described in previous sections (Lee et al., 2010; Tiedemann, 2012).

Additionally, many of the challenges in aligning English to Swedish described by Ahrenberg (2010) would appear to be mitigated.

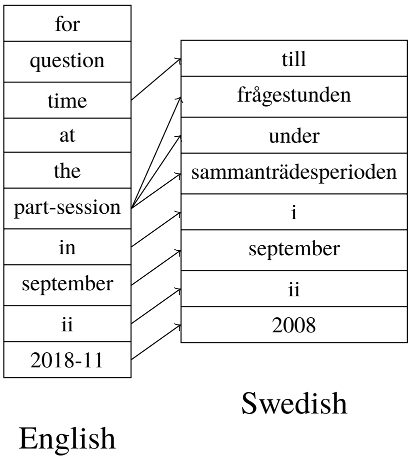

Figure 2.2: Additional example of incorrect projection

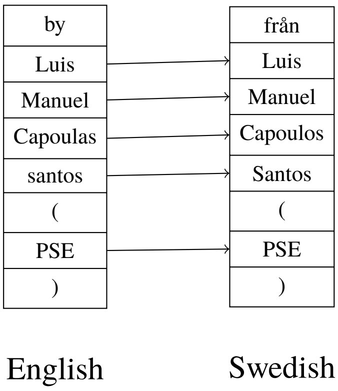

Figure 2.3: Alternative projection

2.4 Bilingual Mention Alignment

Using the word alignments and machine annotation described in Sect. 2.3 the penultimate step to create a training corpora for Swedish or German consists of aligning the coreferring English mentions with mentions in the target language. This was done using a variation of the maximum span heuristic described in Sect. 2.4.1. The heuristic used in this thesis is described in Sect. 2.4.2. The final step was pruning according to a simple percentile heuristic, described in Sect. 2.5.

2.4.1 Maximal Span Heuristic

Bilingual word alignment is complicated even under ideal circumstances as modeling errors, language differences and slight differences in meaning may all affect the word alignment negatively. The Europarl corpus offers additional challenges as the BLEU scores indicate that the sentences seldom correspond to the same meaning - and sometimes are completely disconnected, see Sect. 2.1.

Yarowsky et al. (2001) notes two examples of good and bad projection, reproduced in Figs. 2.4 and 2.5. The figures describe two projection scenarios with varying levels of complexity from a source language on the top of the figures to a target language at the bottom of the figures. The solid lines correspond to word alignments while the dotted lines define the boundaries of their maximum span heuristic. Yarowsky et al. (2001) argues that even though individual word alignments are incorrect that a group of word alignments corresponding to a noun phrase in the source language tend to be grouped together in both the source and target languages. The largest span of aligned words from a noun phrase in the target language usually corresponds to the original noun phrase in the source language.

Using the reasoning of Yarowsky et al. (2001), the maximal span heuristic is to discard any word alignments not mapped to the largest continuous span of the target language and discard overlapping alignments where one mention is not bounded by the other mention for each mention.

The heuristic is non-trivial to evaluate and is primarily selected due to its simplicity and its proven track record with Martins (2015) for the very similar task of creating coreference solvers for South American languages using distant supervision.

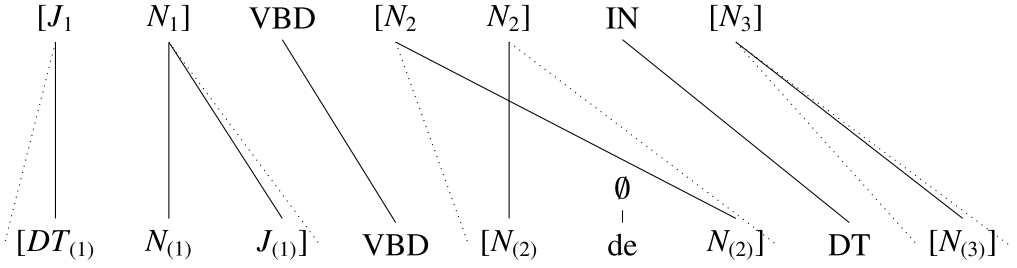

Figure 2.4: Standard projection scenario according to the original paper

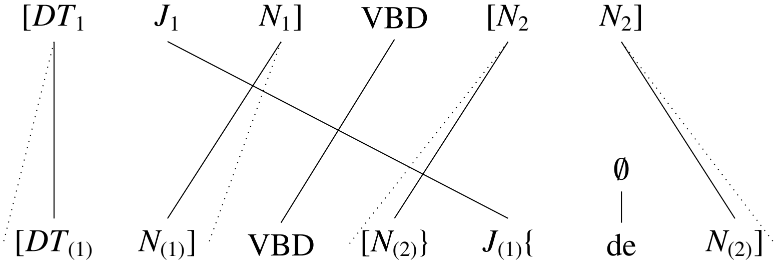

Figure 2.5: Problematic projection scenario according to the original paper

2.4.2 Maximum Span Optimal Mention

The maximum span heuristic described in Sect. 2.4.1 presumes no syntactic knowledge other than tokenization for the target language, which does not factor into the presumptions for this project.

We therefore propose an alteration to the maximum span heuristic which utilizes syntactic knowledge from the target language. The proposed change is to select the largest mention bounded by each maximum span instead of the maximum span itself. This is intuitively better for coreference resolution as the generated corpus would only consist of valid mentions rather than brackets of text without any relation to a mention. This has the additional benefit of simplifying overlapping spans as a mention has a unique head, in accordance with Sect. 5.2, and the problem of overlapping is replaced with pruning mentions with identical bounds.

As with the maximum span heuristic this heuristic is selected by intuition rather than by evaluation as that task would have been non-trivial and outside of the scope for this project.

2.5 Document Pruning

We removed some documents in the corpus from the generated data set according to two metrics; document length and alignment agreement. We reasoned that this would improve on the results out of two considerations; document correctness and mitigation of some modeling constraints in CoreNLP. The reasoning behind removing documents which aligns poorly follows the example of Martins (2015), where documents with poor alignment scores were removed. Misaligned mentions stems from a large number of possible error sources such as poor word alignment and incorrectly annotated parts of speech tags. If a mention missaligns the training set will be affected, which is undesirable. It is warranted to expect some misalignments, but a large proportion of them affects the consistency and validity of the training set and the document in question should be removed. The reasoning behind removing large documents follows from a modelling aspects of CoreNLP, where distant mentions are only considered coreferring if their texts matches perfectly, whereas mentions that are close uses more advanced methods for identifying coreference. We made the assumption that shorter documents would be more likely to have fewer coreference chains containing mentions that were only aligned using exact string match, which is undesirable. Additionally, shorter documents were noticeably faster to annotate which was an unforeseen benefit of only using shorter documents.

The goal was to create a training set with comparable size to the CoNLL task, i.e. a million words or more. To this effect all documents were aligned using our maximum span variant and the alignment accuracy was measured, that is the number of accepted alignments divided by the sum of all alignments.

All documents with lower than average alignment accuracy were removed. Additionally, larger documents were removed until a total training set consisting of approximately a million words in total was generated.

Chapter 3
Data Structures

To support coreference annotation we extended CoNLL-X with a column for coreference using the coreference annotation from the CoNLL 2011 shared task. The format is extensively described in Sect. 3.1. In addition, this section contains a basic description of for dependency grammar with a short example.

3.1 CoNLL-X Format and extension

The definition for the CoNLL format varies depending on the shared task, but the general format is defined by the CoNLL-X format (Buchholz and Marsi, 2006; Nilsson et al., 2007; Pradhan et al., 2011). The general CoNLL format is a document in UTF-8 were each line in the document corresponds to a word or token in a source document. The end of a sentence in the source document corresponds to an empty line in a CoNLL document.

A line containing a word contains additional information about the word and its context arranged as whitespace-delineated columns where each column index contains the information specified in Table 3.1.

The original format is CoNLL-X specified by the CoNLL 2006 shared task (Buchholz and Marsi, 2006) with the addition of the *coreference*-column described in the CoNLL 2011 shared task (Pradhan et al., 2011).

An alternative format would have been to use the CoNLL 2011 format as it is, but we decided against it. The format contains more columns than the CoNLL-X format with data which would be difficult to generate for this project. Specifically the *word sense* and *named entities* columns, which has some German and Swedish analogues, was considered to be too difficult to implement for this project (Hamp et al., 1997; Henrich and Hinrichs, 2010).

Table 3.1: Format Definition

Column Index	Name	Description
1	ID	Ordinal number of the token.
2	FORM	The token with the case and conjugate form of the source document.
3	LEMMA	The token in its base form, if available, and with appropriate upper- and lower case.
3	CPOSTAG	The coarse-grained part of speech form of the token. I.e., the simpler format of the part of speech tag.
4	POSTAG	The token's fine grained part of speech form, equivalent to CPOSTAG if unspecificed.
5	FEATS	An unordered set of the word's syntactic and morphological features separated by a bar (|) if the token has more than one element in the set.
6	HEAD	The head of the current token in regards to dependency parsing.
7	DEPREL	The dependency parser relation between the word and its head as defined by the HEAD column
8	PHEAD	The projective head of the token.
9	PDEPREL	The relation between the word and its projective head.
10	Coreference	Coreference chain information encoded in a parenthesis structure. Delineated by a bar (|) if two or more mentions overlap on that token.

3.2 Dependency Grammar

Dependency grammar views a sentence or a sentence fragment as a projective graph with a root node leading into the phrase's head word (Nivre, 2006). A projective graph means that it is acyclic and that every vertex has at most one incoming edge.

The notion of dependency grammar has its history in descriptive linguistics, which is outside the scope of this thesis (Nivre, 2005).

Figure 3.1: A basic dependency tree

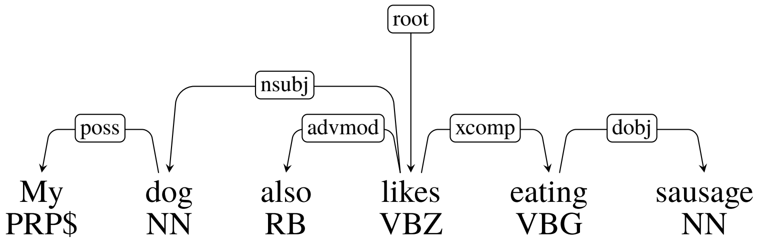

Fig. 3.1 shows a small phrase with part of speech tagging and dependency graph printed

out. As can be seen in Fig. 3.1 the most important word *likes*; the sentence describes something that likes something else.

From the word *likes* there are two outgoing edges; one marked with *nsubj* and one marked with *xcomp*. This means that *dog* is the nominal subject – one who does something and *eating* is the open clause complement (De Marneffe and Manning, 2008)

The graph helps to formally decipher a text and answer questions such as *whose dog?* mine, *what does the dog like?* eating sausages and so forth.

The formalism allows for easier computer analysis of the text content.

3.3 Part-of-Speech Tags

The 24 Swedish part of speech-tags produced by the machine annotator are described in Table 3.2 and they are colloquially known together as a "tag-set". The abbreviations corresponds to a word class in the tagset's language, which usually contains unique word classes particular to the grammatical structure of the language. The German part of speech-tags are not reprinted in this manner due to large number of unique tags.

Table 3.2: SuC Part-of-Speech Categories

Part of speech	Explanation	Part of speech	Explanation
AB	Adverb	PM	Proper Noun
DT	Determiner	PN	Pronoun
HA	WH-adverb	PP	Preposition
HD	WH-determiner	PS	Possessive pronoun
HP	WH-pronoun	RG	Cardinal number
HS	WH-possessive	RO	Ordinal number
IE	Infinitival marker	SN	Subordinating conjunction
JJ	Adjective	VB	Verb
KN	Coordinating conjunction	UO	Foreign word
NN	Noun	MAD	Major delimiter
PC	Participle	MID	Minor delimiter
PL	Particle	PAD	Pairwise delimiter

3.4 Example text

In Table 3.3 the Swedish sentences

> Min hund gillar att äta glass men han gillar också att äta korv. Den stora fluffsiga hunden gör mig galen.

has been annotated with the data described in Fig. 3.1 using machine annotation.
Of particular note is the detected mentions, which can be viewed in the last column, where the mentions *Min hund*, *han* as well as *den stora fluffsiga hunden* are identified as coreferent mentions. Additionally, the mentions *min* and *mig* are identified as mentions as well.

Table 3.3: A basic example text in Swedish

ID	FORM	LEMMA	CPOSTAG	POSTAG	FEATS	HEAD	DEPREL	PHEAD	PDEPREL	Coreference
1	Min	min	PS	PS	UTR|SIN|DEF	2	DT	_	_	(1|(2)
2	hund	hund	NN	NN	UTR|SIN|IND|NOM	3	SS	_	_	1)
3	gillar	gilla	VB	VB	INF|AKT	4	IF	_	_	_
4	att	att	IE	IE	_	3	00	_	_	_
5	äta	äta	VB	VB	INF|AKT	4	IF	_	_	_
6	glass	glass	NN	NN	UTR|SIN|ID|NOM	5	00	_	_	_
7	men	men	KN	KN	_	0	ROOT	_	_	_
8	han	han	PN	PN	UTR|SIN|DEF|SUB	9	SS	_	_	(1)
9	gillar	gilla	VB	VB	PRS|AKT	7	MS	_	_	_
10	också	också	AB	AB	_	9	+A	_	_	_
11	att	att	IE	IE	_	9	00	_	_	_
12	äta	äta	VB	VB	INF|AKT	11	IF	_	_	_
13	korv	korv	NN	NN	UTR|SIN|IND|NOM	12	00	_	_	_
14	.	.	MAD	MAD	_	7	IP	_	_	_
1	Den	den	DT	DT	UTR|SIN|DEF	3	DT	_	_	(1
2	stora	stor	JJ	JJ	POS|UTR/NEU|SIN|DEF|NOM	3	AT	_	_	_
3	fluffsiga	fluffsiga	LE	LE	_	5	00	_	_	_
4	hunden	hund	NN	NN	POS|SIN|DEF|NOM	3	PA	_	_	1)
5	gör	göra	VB	VB	PRS|AKT	0	ROOT	_	_	_
6	mig	jag	PN	PN	UTR|SIN|DEF|OBJ	5	00	_	_	(2)
7	galen	galen	JJ	JJ	POS|UTR|SIN|IND|NOM	5	OP	_	_	_
8	.	.	MAD	MAD	_	5	IP	_	_	_

Chapter 4
Evaluation

This chapter discusses the metrics, vocabulary, and methodology for evaluating a coreference solver for Swedish and German. We use a language-agnostic approach for evaluation with a proven track record in academia.

4.1 Precision, Recall, and other Measurements

The duality of *precision* and *recall* is important to conceptualize and is therefore described in greater detail.

This section make heavy use of a simple fishing allegory to contextualize these concepts. The allegory presents the reader as a fisherman who goes to a lake to catch fish, but who sometimes catches shoes instead. There are people who would prefer shoes over fish, but in the confines of this allegory everyone prefers fish over shoes.

Precision and recall are common evaluation measurements for machine-learning and its application in the field of Natural Language Processing (Powers, 2011). The variables used for the equations in Eq. 4.1, Eq. 4.2 as well as Table 4.2 are explained in Table 4.1.

Precision measures whether a positively identified entity is correctly classified. A good fishing analogue for precision is

> what is the probability that you got a fish and not a shoe?

High precision would mean that one would go home to cook fish, not to open up a shoe shop. The equation for calculating precision is presented in Eq. 4.1.

$$Precision = \frac{T_P}{T_P + F_P} \tag{4.1}$$

Table 4.1: Terminology

Variable	Explanation
T_P	True Positive; correctly classified positive entity *Fish caught*
T_N	True Negative; correctly classified negative entity *Shoes in the lake*
F_P	False Positive; incorrectly classified positive entity *Shoes caught*
F_N	False Negative; incorrectly classified negative entity *Fish in the lake*

Recall measures the probability that positive entities are correctly classified as positive. A good fishing analogue for recall is *"How large is the fishing net?"*. A low recall value would mean that one brought home less fish (or shoes). The equation for calculating recall is presented in Eq. 4.2.

$$Recall = \frac{T_P}{T_P + F_N} \tag{4.2}$$

If one were to catch only a single fish (and no shoes) and then go home the precision value would be extremely high, though the recall values would be horrendous (There is still a lot of fish in the lake). Alternatively, if one were to bring home everything in the lake (high recall) would also have to contend with a lot of shoes (low precision).

Ideally, a well performing system need to balance the two extremes, i.e. bring home as much fish as possible but try to keep the shoes to a minimum.

The two measurements are commonly presented in the form of the F1-score, which calculates the harmonic mean of the precision and recall values, see Eq. 4.3.

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.3}$$

Both precision and recall are classification measurements, which is a kind of metric for when one wants to know if an entity belongs to a group. Neither precision nor recall measures true negatives, which means that neither is interested in how many shoes a lake has – only fish (or at least what passes as fish).

Confusion matrices, such as the matrix in Table 4.2, offers an intuitive representation of a data set, where both precision and recall may be read intuitively in addition to the size and magnitude of the corpus as a whole.

Table 4.2: Basic confusion matrix

		Predicted		
		Positive	Negative	Total
Actual	Positive	T_P	F_N	$T_P + F_N$
	Negative	F_P	T_N	$F_P + T_N$
	Total	$T_P + F_P$	$F_N + T_N$	$T_P + F_N + F_P + T_N$

4.2 Metrics

This section specifies the metrics considered by the CoNLL 2012 shared task. CoNLL 2012 was particularly interested in a single value metric to evaluate and compare coreference solvers with different performance characteristics fairly and decided on the $MELA_{CoNLL}$ metric, described in more detail in Sect. 4.2.5 (Pradhan et al., 2012).

4.2.1 $MUC6$

$MUC6$ is the oldest metric considered for the CoNLL 2011 and 2012 shared tasks. The metric was originally proposed by Vilain et al. (1995) and models the links between mentions as vertices in a graph and the metric is a measurement of how many links should be inserted or deleted to create a spanning tree, i.e. how many mentions must be to be added or removed from coreference chains so that all chains are correct. Cai and Strube (2010) argues that the metric measures singletons incorrectly and would therefore be a poor choice for a real world situation (Cai and Strube, 2010).

4.2.2 B^3

Bagga and Baldwin (1998) proposed the B^3 as an improvement upon $MUC6$. The link-based algorithm calculates precision and recall for each mention including singletons. Stoyanov et al. (2009) note that B^3 assumes that the gold standard and the system response clusters over the same set of mentions, which is incorrect for the case when the coreference solver uses a separate method to identify mention, i.e. the metric does not sufficiently penalize clustered mentions that were not part of the test set.

4.2.3 $CEAF$

$CEAF_E$ is a entity based metric proposed by Luo (2005) that attempt to map the system response entities to the entities in the gold standard, keeping the best and discarding the rest. Stoyanov et al. (2009) note that this approach exhibits low precision when the solver uses a separate method to identify entities as it marks each singleton outside of the gold standard as zero and takes the average of all chains independent of size.

$CEAF_M$ uses mentions rather than entities, but $CEAF_E$ is the metric used in the CoNLL 2012 shared task (Pradhan et al., 2012).

4.2.4 *BLANC*

Recasens and Hovy (2011) proposed using a variation of the Rand Index (Rand, 1971), which measures the similarity between two clusters. As with other metrics *BLANC* was proposed to resolve issues with previously proposed metrics. As with previous metrics *BLANC* has weaknesses, amongst other the problem when measuring documents with few coreferring mentions where small errors will affect the final score greatly (Stamborg, 2012).

4.2.5 $MELA_{CoNLL}$

As the previously proposed metrics has both benefits and flaws the CoNLL 2012 uses the $MELA$ metric proposed by Denis et al. (2009) which takes the weighted average of $MUC6$, B^3 and $CEAF_E$, i.e., the MELA equation in Eq. 4.4. The proposed benefits lies in that each metric represents an important dimension; $MUC6$ is based on links, B^3 is based on mentions and $CEAF_E$ is based on entities. As the CoNLL 2012 task decided to use the unweighted mean of the three metrics, i.e., in accordance with Eq. 4.5 (Pradhan et al., 2012).

$$MELA = \frac{MUC6}{a} + \frac{B^3}{b} + \frac{CEAF_E}{c} \tag{4.4}$$

$$MELA_{CoNLL} = \frac{MUC6 + B^3 + CEAF_E}{3} \tag{4.5}$$

4.3 SUC-Core (Swedish)

The SUC-Core corpus consists of 20,000 words and tokens in 10 documents with 2,758 coreferring mentions created by Kristina Nilsson Björkenstam at Stockholm University (Nilsson Björkenstam, 2013). The corpus is a subset of the SUC 2.0 corpus, annotated with noun phrase coreferential links (Gustafson-Capková and Hartmann, 2006).

The corpus is much too small to train a well rounded coreference solver, but the corpus is more than sufficient to evaluate solvers trained on some different source material.

The corpus is supplied in a non-standardized format without part of speech or dependency graph information. Both part of speech and dependency graph data is readily available in the SUC 2.0 and SUC 3.0 corpora.

As a preparatory step to evaluate coreference resolution in Swedish the information from SUC-Core was merged with SUC 2.0 and SUC 3.0 to a CoNLL 2012 compatible file format and redistributed to the corpus' original creator. Additionally, singletons were removed from the merged data files.

4.4 Tüba D/Z (German)

The Tüba D/Z corpus consists of 1,787,801 words and tokens organized in 3,644 files annotated with both part of speech and dependency graph information.

Tüba D/Z was created by the department of philosophy at the University of Tübingen, Germany (Henrich and Hinrichs, 2013)(Henrich and Hinrichs, 2014). Although the corpus would be sufficient in size to train a coreference solver it is only used for evaluation in this thesis. As with SUC-Core all singletons were removed from this corpus prior to being used for evaluation. Due to time and memory constraints only a small subset of the Tüba D/Z corpus was used for evaluation.

4.5 End-to-End Evaluation

Similarly to the CoNLL 2011 and 2012 shared tasks, we evaluated our system using gold and predicted mention boundaries. When given the gold mentions, the solver knows the boundaries of all nonsingleton mentions in the test set, while with predicted mention boundaries, the solver has no prior knowledge about the test set. We also followed the shared tasks in only using machine-annotated parses as input.

The rationale for using gold mention boundaries is that they correspond to the use of an ideal method for mention identification, where the results are an upper bound for the solver as it does not consider singleton mentions Pradhan et al. (2011).

Finally, the solver's coreference resolution will be evaluated using the metrics described in Sect. 4.2, i.e. "The CoNLL-metrics".

Chapter 5
System Architecture

There are two common machine-learning approaches for coreference resolution; variations of the closest antecedent approach described in the seminal paper by Soon et al. (2001) and the latent tree approach proposed by dos Santos and Carvalho (2011). The latent tree approach is a later development in the field, but shows much promise. During the CoNLL 2012 shared task a team using latent trees obtained the best results (Pradhan et al., 2012).

After considerations it was decided to create a solver using the closest antecedent approach; the closest antecedent approach is less complicated which would hopefully be sufficiently robust to account for systemic errors from the mention transfer.

5.1 Document Selection

A subset consisting of the shortest documents containing at least one coreference chain were used for training. After selection and pruning the Swedish training set consisted of 4,366,897 words and 183,207 sentences in 1,717 documents.

5.2 Mention Identification

Endocentric and exocentric are linguistic terms that describes how a grammatical construct such as a phrase functions in a text. An endocentric construct has a head element and one or more dependents and functions grammatically as the head element's word class. An exocentric construct consists of two or more parts where neither part describes the grammatical function of the whole construct (Bloomfield, 1935). Dependency grammar is by definition endocentric and lacks the formal phrase notation of constituent grammars (Nivre, 2005).

A noun phrase is a linguistic term for an *endocentric* or *exocentric* construct that fulfills the linguistic function of a noun (Bloomfield, 1935). Noun phrases are a well established concept in many languages such as English (Huddleston et al., 2002), German (Lehmann, 1957) and Swedish (Hultman, 2003) today, but the origins are somewhat obscure and may date back to as late as Leonard Bloomfield in the beginning of the 19th century (Hudson, 1994).

As Malt Parser only offers dependency grammar models for Swedish some additional modeling was required to identify noun phrases (Nivre et al., 2007).

By analyzing the noun phrases in SUC-Core some basic patterns could be discerned; most noun phrases were bounded by a subset of the dependency tree with the dominant word at its head. The head word commonly belonged to a small subset of the part-of-speech tag set. Hand-written rules were written to approximate noun phrase identification using these basic observations. The identified noun phrases were then post-processed with additional rules to better map their boundaries with the phrase boundaries in SUC-Core.

The handwritten rules for identifying noun phrases using dependency grammar can be found in Table 5.1. Rules further up takes precedence over rules further down in the table.

The identified noun phrases were then post-processed to better align with the boundaries defined by SUC-Core according to the rules in Table 5.2.

Table 5.1: Hand-written rules for noun phrase identification for Swedish based on SUC-Core

Head POS	Additional rule	NP
UO	Dependency head has POS PM	No
	Otherwise	Yes
PM	Dependency head has POS PM but different grammatical case	Yes
	Dependency head has POS PM	No
	Otherwise	Yes
PS		Yes
PN		Yes
NN		Yes
KN	the head word is “och” and has at least one child who is a mention	Yes
	Otherwise	No
DT	the head word is “den”	Yes
	Otherwise	No
JJ	the head word is “själv”	Yes
	Otherwise	No

Table 5.2: Additional hand-written rules for post processing identified noun phrases

Ordinal number	Rule
1	Remove words from the beginning or the end of the phrase if they have the POS tags ET, EF or VB.
2	The first and last words closest to the mention head with the HP POS tag and all words further from the mention head is removed from the phrase.
3	Remove words from the beginning or the end of the phrase if they have the POS tags AB or MAD.
4	The first and last words closest to the mention head with the HP POS tag and with the dependency arch SS and all words further from the mention head is removed from the phrase.
5	Remove words from the end of the phrase if they have the POS tag PP.
6	Remove words from the beginning or the end of the phrase if they have the POS tag PAD.

5.3 Algorithms

This section describes the algorithms used for solving coreference. How the training set is generated and how classifiers are applied to mentions is described in exhaustive detail, while the machine-learning aspects are simply noted.

5.3.1 Generating the Training Set Using Closest Antecedent

The closest antecedent approach models coreference chains as a series of linked mentions, where every mention has at most one antecedent and one anaphora. The modeling assumptions relaxes the complex relationship between coreferring mentions by only considering the relationship between a mention and its closest antecedent.

As only the closest antecedent is considered when solving a document's coreference the problem is reconstituted into a binary classification problem where the system only need to consider whether a mention and its closest antecedent corefers (Soon et al., 2001).

Figure 5.1 shows a series of coreferring mentions, which could be found in a typical document. The actual text is inconsequential and only the mentions are presented. The figure presents the coreferring mention chains as A, B and C with an ordinal number unique for each chain. The order of the mentions in the figure represents their natural order in which they would appear in a document.

The positive training set is generated by mapping each coreferring mention with its closest preceding mention with which it coreffers. Each such positive mention pair is considered a positive sample from the training set. The mentions in Fig. 5.1 are graphically presented in Fig. 5.2. The negative training set is generated by identifying every intervening mention between each positive mention pair and consider its relation with the positive mention pair's anaphora. The graphical visualization of this can be viewed in Fig. 5.3. These relations may also be viewed as a table in Table 5.3.

Naturally, only coreferring mentions are considered when generating the training set; mentions identified by the coreference solver are not part of the training set. When generating a training set the negative training samples are often an order of magnitude more numerous than the positive training samples which may skew the model. CoreNLP solves this discrepancy by capping the ratio at 5% for its neural network based coreference solver as default. We randomly select negative training samples until the positive training examples consists of approximately 4 to 5 % of the traning set.

A_1 B_1 C_1 C_2 B_2 C_3 C_4 B_3 A_2

Figure 5.1: Sequence of mentions extracted in order from three coreference chains in a typical document

Table 5.3: The training data generated from Fig. 5.1 as a table

Antecedent	Anaphora	Type
A_1	A_2	Positive
B_1	B_2	Positive
C_1	C_2	Positive
C_2	C_3	Positive
C_3	C_4	Positive
B_2	B_3	Positive
B_1	A_2	Negative
C_1	A_2	Negative
C_2	A_2	Negative
C_1	B_2	Negative
C_2	B_2	Negative
B_2	A_2	Negative
B_2	C_3	Negative
C_3	B_3	Negative
C_3	A_2	Negative
C_4	B_3	Negative
C_4	A_2	Negative
B_3	A_2	Negative

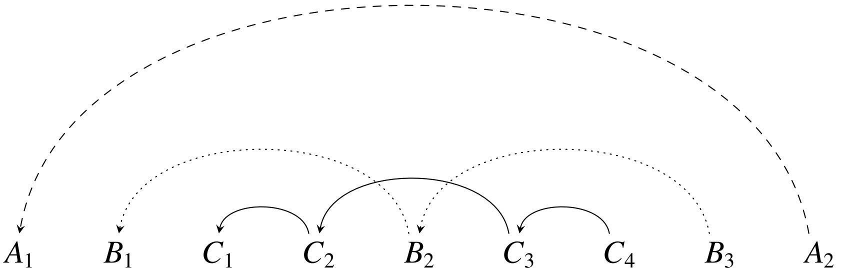

Figure 5.2: Positive training examples. Line pattern according to which coreference chain the anaphora mention belongs to

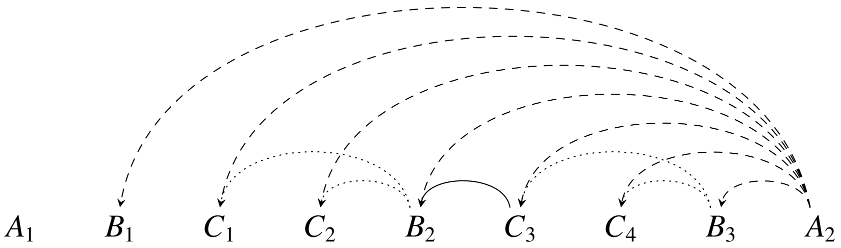

Figure 5.3: Negative training examples. Line pattern according to which coreference chain the anaphora mention belongs to

5.3.2 Machine-Learning Algorithms

The Weka Toolkit and LibLinear was used for the machine-learning aspects of the coreference solver. The C4.5, Random Forest and Logistic Regression algorithms were used for the binary classification problem (Witten and Frank, 2005; Hall et al., 2009; Fan et al., 2008).

5.3.3 Solving Coreference

Corefering chains are identified using the methods described by Soon et al. (2001). The text in the section describes this procedure.

Mentions are identified according to Sect. 5.2 and ordered according to their position relative to the start of the document.

Coreference chains are identified by creating the empty set C and iterating over every mention in order from the last mention in the document to the first. If the mention corefers with the closest mention in the C set or if the set is empty the mention is added to the set. Once every mention has been iterated over, the iteration ends. If the set contains more than one mention it is a coreferring chain. As long as there are mentions which has not been added to a set the a new loop is initiated, iterating over the mentions that has not yet been added to a set. The procedure is formulated in pseudo-code in Fig. 5.4.

```
procedure Closest-Antecedent
    $M \leftarrow allMentions()$
    $C \leftarrow \emptyset$
    while $M \neq \emptyset$ do
        $C_n \leftarrow \emptyset$
        for $m \leftarrow M_{last}$ to $M_{first}$ do
            if $C = \emptyset \cup$ corefers($m$, $C_{closest}$) then
                $C_n = C_n \cup \{m\}$                ▷ $m$ is added to $C_n$
                $M = M \setminus m$                   ▷ m is removed from M
            end if
        end for
        if size($C_n$) > 1 then
            $C = C \cup C_n$                          ▷ $C_n$ is added to $C$ as a chain
        end if
    end while
    return $C$
end procedure
```

Figure 5.4: Pseudo-code for solving a documents coreference using closest antecedent

5.4 Features

Coreference resolution can be modelled as a binary classification problem, as described in Sect. 5.3. In this approach the classifier needs a set of features to distinguish coreferring from nonreferring mention pairs.

The feature set is dependent on what information can be extracted from the text - lexical features such as parts-of-speech, classification of proper names as well as gender disambiguation for personal names. Models using large amounts of contextual information in its feature sets generally performs better than models with smaller feature sets. Training data is often contradictory, especially when modelling human text, which dictates that models should only be as complicated as the prerequisites requires to limit the risk of overfitting the data. Overfitting means that the model describes random noise rather than actual relationships in the data, which affects the model's performance negatively. As we use distant supervision it is likely that our generated training data will be noisy, so we decided to use a small feature set to evaluate the validity of our approach rather than create a fully featured coreference solver.

The feature set is described in Tab. 5.4 and is a subset of the feature sets of Stamborg et al. (2012) and Soon et al. (2001).

Additional features such as named entity recognition would be possible to integrate into our model using Stagger (Östling, 2013), but we decided to only use the syntactic information from the data described in Sect. 3.1 for the sake of simplicity.

Axelsson et al. (2014) uses a hard-coded distance feature. This feature only considers mentions with less than 160 intervening mentions when modeling coreference. Even though our mentions are identified using another approach we decided to use the same heuristic.

Table 5.4: The feature set used for Swedish

Rule	**Description**	**Type**
StrEquivalence	Mentions are identical	Boolean
HeadStrEquivalence	Mention head words are identical	Boolean
AnaphoraIsPN	POS of anaphora head word is PN	Boolean
AntecedentIsPN	POS of antecedent head word is PN	Boolean
AnaphoraIsPM	POS of anaphora head word is PM	Boolean
AntecedentIsPM	POS of antecedent head word is PM	Boolean
AnaphoraHasDT	Anaphora head word has the morphological feat DT	Boolean
AntecedentArticle	Antecedent head grammatical article	Enum
AnaphoraArticle	Anaphora head grammatical article	Enum
AntecedentGrammaticalNumber	Antecedent grammatical number	Enum
AnaphoraGrammaticalNumber	Anaphora grammatical number	Enum
AntecedentGender	Antecedent grammatical gender	Enum
AnaphoraGender	Anaphora grammatical gender	Enum

Chapter 6
German Modeling

Modelling for German coreference closely follows the system architecture described in Chapter 5, with some slight alterations to account for the linguistic differences. This section describes the adjustments necessitated to create a German model based on the approach developed for the Swedish language. This chapter only contains the adjustments; if a section is absent in this chapter it only means that the methods described for Swedish corefence did not need adjustments to model German coreference.

6.1 Document Selection

A subset consisting of the randomly selected documents containing at least one coreference chain were used for training. After selection and pruning the training set consists of 9,028,208 words and 342,852 sentences in 1,717 documents.

6.2 Mention Identification

The German models for Mate Tools are trained on the TIGER corpus, which was used for the CoNLL 2009 shared task and is one of the standard treebanks for the German language (Hajič et al., 2009). The Tüba D/Z corpus was used for rule adaptation for German.

Identifying mentions using the same approach as described in Sect. 5.2 as more post processing was required to yield comparable results.

Noun phrase identification in German proved more complicated than noun phrase identification in Swedish. One example of this is the identification of split antecedents. Consider the phrase *Anna and Paul*. *Anna* and *Paul* are possible mentions, but so is the whole phrase. In Swedish the corresponding phrase would be *Anna och Paul* with the conjunction *och* as the head word. The annotation scheme used for the TIGER corpus does not have the conjunction as head for coordinated noun phrases (Albert et al., 2003). In Swedish the rule for identifying split antecedents only needed to check whether a conjunction had children that were noun phrases, whereas in German the same rule required more analysis.

Identifying split antecedents in German is primarily an example of the inherit challenges of analysing languages without sufficient linguistic proficiency in a particular language; the amount of contextual information is limited.

Given some additional post processing steps compared with Swedish the hand-written rules for noun phrase identification in German can be seen in Fig. 6.1 with the post-processing rules in Fig. 6.2.

Table 6.1: Hand-written rules for noun phrase identification for German based on Tüba-D/Z

Head POS	**Additional rule**	**NP**
NN	Dependency head has POS NN	No
	Otherwise	Yes
NE	Dependency head has POS NE	No
	Otherwise	Yes
PRELS		Yes
PRF		Yes
PPOSAT		Yes
PRELAT		Yes
PIS		Yes
PDAT		Yes
PDS		Yes
FM		Yes
CARD		Yes

Table 6.2: Additional hand-written rules for post processing identified noun phrases in German

Ordinal number	**Rule**
1	Remove words from the start or the end of the phrase if they have the POS tags $. $(PROP KON.
2	If there is a word with the POS tag VVPP after the head word the word prior to this word becomes the last word in the phrase.
3	If there is a dependant word with the POS tag KON and its string equals *und* create additional mentions from the phrases left and right of this word.
4	If there is a word with the POS tag APPRART after the head word the word prior to this word becomes the last word in the phrase.

6.3 Feature Set

The feature set for German is described in Tab. 6.3. The primary difference between German and Swedish is the addition of gender classified names and IMS Hotcoref DE contained lists of names and job titles which were applied to the training of the German model.

The morphological information from both CoreNLP and Mate Tools appeared to be limited when compared with Swedish which is reflected in the feature set.

Table 6.3: The feature set used for German

Rule	Description	Type
StrEquivalence	Mentions are identical	Boolean
HeadStrEquivalence	Mention head words are identical	Boolean
AnaphoraIsMale	Checks if mention contains a word which is a male first name	Boolean
AnaphoraIsFemale	Checks if mention contains a word which is a female first name	Boolean
AnaphoraIsPerson	Checks if mention contains a word which is a job title	Boolean
AntecedentIsMale	Checks if mention contains a word which is a male first name	Boolean
AntecedentIsFemale	Checks if mention contains a word which is a female first name	Boolean
AntecedentIsPerson	Checks if mention contains a word which is a job title	Boolean
MentionDistance	Number of intervening sentences between the two mentions. Capped at 10.	Integer
AntecedentGrammaticalGender	Grammatical gender of antecedent head word	Enum
AnaphoraGrammaticalGender	Grammatical gender of anaphora head word	Enum
AnaphoraSubj	Anaphora head is subject	Enum
AntecedentSubj	Antecent head is subject	Enum
AnaphoraGen	Anaphora has the morphological feature gen	Enum
AntecedentGen	Antecedent has the morphological feature gen	Enum
AnaphoraInd	Anaphora has the morphological feature ind	Enum
AntecedentInd	Antecedent has the morphological feature ind	Enum
AnaphoraNom	Anaphora has the morphological feature nom	Enum
AntecedentNom	Antecedent has the morphological feature nom	Enum
AnaphoraSg	Anaphora has the morphological feature sg	Enum
AntecedentSg	Antecedent has the morphological feature sg	Enum

Chapter 7
Results

This chapter describes the results of our experiments ordered in three sections; the creation of the training set using distant supervision is presented in Sect. 7.1, identification of mentions using dependency grammar is presented in Sect. 7.2 and finally the system results for the coreference resolution is presented in Sect. 7.3.

7.1 Mention Alignment

This section describes the results from aligning English mentions with a target language using word alignment and the maximum span optimal span heuristic.

7.1.1 Swedish Mention Alignment

The Swedish EuroParl corpora consists of 8,445 documents. From these documents a subset consisting of 3,445 documents were selected based on size, where smaller documents were preferable.

The selected documents contained in total 1,189,557 mentions that were successfully transferred and 541,608 rejected mentions.

Every document with less than 70% successfully transferred documents were removed, which yielded a final tally of 515,777 successfully transferred mentions and 198,675 rejected mentions in 1,717 documents.

7.1.2 German Mention Alignment

The German EuroParl corpora consists of 8,446 documents. From these documents a subset consisting of 2,568 documents were randomly selected.

The selected documents contained in total 992,734 successfully transferred and 503,690 rejected mentions.

Every documents with less than 60% successfully transferred documents were removed, which yielded a final tally of 975,539 successfully transferred mentions and 491,009 rejected mentions in 964 documents.

7.2 Mention Identification

This section describes how well mentions were identified using the rule based dependency grammar described in previous sections.

7.2.1 Swedish Mention Identification

Using the rules described in Table 5.1 91.35% of mentions were identified in SUC-Core.

With the additional post processing rules described in Table 5.2 the results were improved to 95.82%.

7.2.2 German Mention Identification

Using the rules described in Table 5.1 65.90% of mentions were identified in Tüba D/Z.

With the additional post processing rules described in Table 5.2 the results were improved to 82.08%.

7.3 Coreference Resolution

This section presents the end-to-end results for coreference resolution. In addition, this section also presents the results of the classification problem described in Sect. 5.3.3 using the algorithms mentioned in Sect. 5.3.2.

7.3.1 Mention Classification as Coreferring

This section describes the results of the classification problem described in Sect. 5.3.3. The values from this section derives from applying the trained model to the training data.

Swedish Mention Classification

J48 yields a precision of 90.4% and a recall rate of 31.5%, which yields a F-measure of 46.74. The confusion matrix for J48 is described in Table 7.1.

Random forest yields a precision of 91.4% and a recall rate of 31.64%, which yields a F-measure of 47. The confusion matrix for Random Forest is described in Table 7.2.

Logistic regression yields a precision of 80.72% and a recall rate of 33.76%, which yields a F-measure of 47.6 . The confusion matrix for logistic regression is described in Table 7.3.

Table 7.1: Confusion Matrix for J48 on the Swedish Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	116 297	252 749	369 046
	Negative	12 307	5 840 275	5 852 582
	Total	128 604	6 093 024	6 221 628

Table 7.2: Confusion Matrix for Random Forest on the Swedish Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	116 765	252 281	369 046
	Negative	10 980	5 847 112	5 858 092
	Total	127 745	6 099 393	6 227 138

Table 7.3: Confusion Matrix for Logistic Regression on the Swedish Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	124 601	244 445	369 046
	Negative	29 753	5 822 571	5 852 324
	Total	1 543 54	6 067 016	6 221 370

German Mention Classification

J48 yields a precision of 91.74% and a recall rate of 77.68%, which yields a F-measure of 84.13. The confusion matrix for J48 is described in Table 7.4.

Random forest yields a precision of 92.85% and a recall rate of 79.6%, which yields a F-measure of 85.71. The confusion matrix for random forest is described in Table 7.5.

Logistic regression yields a precision of 91.47% and a recall rate of 77.02%, which yields a F-measure of 83.62. The confusion matrix for logistic regression is described in Table 7.6.

Table 7.4: Confusion Matrix for J48 on the German Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	560 611	161 054	721 665
	Negative	50 481	4 675 578	4 726 059
	Total	611 092	4 836 632	5 447 624

Table 7.5: Confusion Matrix for Random Forest on the German Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	574 435	147 230	721 665
	Negative	44 199	4 680 357	4 724 556
	Total	618 634	4 827 587	5 446 223

Table 7.6: Confusion Matrix for Logistic Regression on the German Training Set

		Predicted		
		Positive	Negative	Total
Actual	Positive	555 795	165 870	721 665
	Negative	51 847	4 677 297	4 729 144
	Total	607 642	4 843 167	5 450 809

7.3.2 End-to-End Results

The results are presented in two tables, Table 7.7 shows the results using predicted mentions and Table 7.8 shows the results using gold mentions.

Table 7.7: End-to-end results using predicted mentions

Language	**method**	$MUC6$	B^3	$CEAF_E$	$CEAF_M$	$BLANC$	$MELA_{CoNLL}$
Swedish	J48	**46.72**	**29.11**	**28.32**	**32.67**	**29.94**	**34.98**
	Random Forest	46.29	28.87	27.68	32.21	29.41	34.28
	Logistic Regression	39.18	2.4	1.01	8.88	5.46	14.19
German	J48	**34.29**	**2.63**	**2.55**	**12.81**	4.67	**13.16**
	Random Forest	33.51	2.54	2.4	11.82	**5.46**	12.81
	Logistic Regression	33.97	2.36	1.35	12.5	4.58	12.56

Table 7.8: End-to-end results using gold mentions

Language	**method**	$MUC6$	B^3	$CEAF_E$	$CEAF_M$	$BLANC$	$MELA_{CoNLL}$
Swedish	J48	61.43	**37.78**	40.97	42.36	**41.51**	**46.73**
	Random Forest	61.37	37.72	**41.03**	**42.46**	41.22	46.71
	Logistic Regression	**84.77**	13.37	1.95	16.68	15.5	33.37
German	J48	82.69	19.74	5.86	26.75	19.56	36.1
	Random Forest	77.24	**24.16**	**9.53**	**26.94**	**32.72**	**36.98**
	Logistic Regression	**83.71**	17.6	4.5	25.58	16.61	35.27

Chapter 8
Discussion

8.1 Discussion

The primary challenge for this Master's thesis was to create a system that would give adequate performance end-to-end when solving coreference; at the onset of the thesis work the lack of an existing coreference solver, mention identification and a gold corpus in a CoNLL compatible format greatly affected our ability to evaluate our progress. All of these obstacles were cleared at various steps along the way with good help from various people.

The lack of commonly available tools to support coreference resolution in Swedish indicates to us that the challenges for creating coreference solvers in new languages lies not only in the creation of a suitable training corpus, but also with evaluation and other subtasks. The CoNLL 2011 shared task offers both the tools and the metrics to evaluate coreference, but without good mention identification and a reference corpora with solved coreference resolution these tools are of little consequence.

The feature sets described in previous sections were limited to simple linguistic features defined by the annotation schema defined by the annotation schema for the Swedish treebank and the annotation schema for the German Tiger corpus respectively. A large subset of the feature set of Stamborg et al. (2012) would very likely improve on the results in this thesis, but for reasons described in previous sections they were not added. Some features such as named entity recognition and knowledge bases such as WordNet is available for both Swedish and German, but neither were used for this project due to time and resource constraints.

The results in Tables 7.7 and 7.8 show that even though the dependency grammar based approach for identifying probable mentions gives adequate performance, a stronger focus on identifying and pruning mentions would greatly improve on the results. The gold mentions in Table 7.8 are unrealistic for a real life scenario as the gold mentions lacks singletons, but should rather be considered as the best result reachable by the solver given an ideal mention identification. The great improvements for especially German between the results in Tables 7.7 and 7.8 indicates that the primary difference between Swedish and German lies in the quality of the mention identification rather than the feature set of the coreference solver.

The results in Sect. 7.3.1 presents J48, random forest and logistic regression as comparable classifiers for the data sets, but this is not the case in Sect. 7.3.2. When the system ran end-to-end using logistic regression the end results performed worse than the other algorithms for most metrics, which goes to show that measuring small aspects of a system is not by necessity a good indicator of its performance and that the best evaluation is a complete system measured end-to-end.

8.2 Conclusions

Coreference resolution using weak labelled training data from distant supervision shows great promise for improving coreference resolution for Swedish and other languages. Despite a noisy multilingual corpora were used for mention alignment the generated training corpora was sufficient to train a competent coreference solver.

Coreference resolution improves greatly when gold mentions are used, which implies that improved mention detection is a prerequisite for good results in coreference resolution.

8.3 Future Work

This Master's thesis explores distant supervision as a possible step towards solving coreference in Swedish and German. The approach is general and would benefit from additional attention to various subtasks. A short list is enumerated below.

Using latent trees for improving the coreference resolution. The current state of the art uses latent trees for solving coreference rather than the closest antecedent approach used in this Master's thesis. Commonly available solvers that solve coreference using latent trees appear to use constituent parse trees, which would be a challenge when applied to a language such as Swedish which lacks constituent parse tree models.

Improving and evaluating methods for mention alignment. Martins (2015) uses a far higher threshold for mention alignment than what was feasible for this project. There are many possible reasons for this such as different alignment methods and multilingual corpora that are better aligned, which calls for further study as improper mention alignments have a huge potential impact on the final results.

Improving on mention identification. The current approach of using hand-written rules for mention identification is sufficient for Swedish, but less than ideal for German. There are many possible reasons for this such as language proficiency and the methods used for identifying mentions for the hand annotated corpora. A machine-learning approach such as neural networks offers a better approach which could offer improved performance while requiring less language proficiency.

Evaluating coreference without hand annotated corpora. Multiple languages lacks even the smallest hand annotated corpus for evaluating coreference. A possible approach would be to use the methods in this Master's thesis in reverse and evaluate the aligned mentions. If a monolingual subset of a multilingual corpora would be hand annotated it would allow evaluation of the other languages in the corpus. Experience from working with the Europarl corpus tells us that sentences aligns differently depending on which language pair is used, which would be the fundamental challenge with this approach.

Bibliography

Ahrenberg, L. (2010). Alignment-based profiling of Europarl data in an English-Swedish parallel corpus. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Al-onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, Final Report, JHU Summer Workshop.

Albert, S., Anderssen, J., Bader, R., Becker, S., Bracht, T., Brants, S., Brants, T., Demberg, V., Dipper, S., Eisenberg, P., et al. (2003). Tiger annotationsschema. Technical report, Universität des Saarlandes.

Axelsson, F., Rydback, B., Johansson, F., Bengtsson, J., and Marinov, S. (2014). *Data-driven Coreference Resolution for Swedish*. PhD thesis, Master's thesis, Chalmers University of Technology.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 79–85, Stroudsburg, PA, USA. Association for Computational Linguistics.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.

Björkelund, A. and Kuhn, J. (2014). Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.

Bloomfield, L. (1935). Language. rev. ed. Page 194 is particularly interesting.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., and Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Research on language and computation*, 2(4):597–620.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Cai, J. and Strube, M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 28–36. Association for Computational Linguistics.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University.

Denis, P., Baldridge, J., et al. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, 42(1):87–96.

dos Santos, C. N. and Carvalho, D. L. (2011). Rule and tree ensembles for unrestricted coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 51–55. Association for Computational Linguistics.

Exner, P., Klang, M., and Nugues, P. (2015). A distant supervision approach to semantic role labeling. In *Fourth Joint Conference on Lexical and Computational Semantics (* SEM 2015)*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. (2010). The ai behind watson—the technical article. *AI Magazine*.

Gustafson-Capková, S. and Hartmann, B. (2006). Manual of the stockholm umeå corpus version 2.0. Technical report, Stockholm University.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., et al. (2009). The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hamp, B., Feldweg, H., et al. (1997). Germanet – a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Workshop Proceedings*.

Henrich, V. and Hinrichs, E. (2010). Gernedit – the germanet editing tool. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Henrich, V. and Hinrichs, E. (2013). Extending the tüba-d/z treebank with germanet sense annotation. In *Language Processing and Knowledge in the Web*, pages 89–96. Springer Berlin Heidelberg.

Henrich, V. and Hinrichs, E. (2014). Consistency of manual sense annotation and integration into the tüba-d/z treebank. In *Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT13)*.

Huddleston, R., Pullum, G. K., et al. (2002). The cambridge grammar of english. *Language. Cambridge: Cambridge University Press*, pages 1–23.

Hudson, R. (1994). Discontinuous phrases in dependency grammar.

Hultman, T. G. (2003). *Svenska akademiens språklära*. Svenska akademien.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(03):311–325.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Lee, J.-H., Lee, S.-W., Hong, G., Hwang, Y.-S., Kim, S.-B., and Rim, H.-C. (2010). A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 623–629. Association for Computational Linguistics.

Lehmann, W. (1957). Structure of noun phrases in german. In *Proceedings of the Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Martins, A. F. T. (2015). Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1427–1437, Beijing, China. Association for Computational Linguistics.

Nilsson, J., Riedel, S., and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932. sn.

Nilsson Björkenstam, K. (2013). Suc-core: A balanced corpus annotated with noun phrase coreference. *Northern European Journal of Language Technology (NEJLT)*, 3:19–39.

Nivre, J. (2005). Dependency grammar and dependency parsing. Technical report, Växjö University: School of Mathematics and Systems Engineering.

Nivre, J. (2006). Dependency parsing. *Inductive Dependency Parsing*, pages 45–86.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics.

Östling, R. (2013). Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27. Association for Computational Linguistics.

Rahman, A. and Ng, V. (2012). Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730. Association for Computational Linguistics.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510.

Rösiger, I. and Kuhn, J. (2016). Ims hotcoref de: A data-driven co-reference resolver for german. In *LREC*.

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

Stamborg, M. (2012). Statistical coreference resolving in a multi-language domain.

Stamborg, M., Medved, D., Exner, P., and Nugues, P. (2012). Using syntactic dependencies to solve coreferences. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 64–70. Association for Computational Linguistics.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. (2009). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 656–664. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yao, L., Riedel, S., and McCallum, A. (2010). Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023. Association for Computational Linguistics.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

EXAMENSARBETE Creating a coreference solver for Swedish and German using distant supervision
STUDENT Alexander Wallin
HANDLEDARE Pierre Nugues (LTH)
EXAMINATOR Jacek Malec (LTH)

Do I refer to you?

A possible step towards solving the riddle of Swedish coreference

POPULÄRVETENSKAPLIG SAMMANFATTNING **Alexander Wallin**

It is said that coreference is difficult to explain, but easy to comprehend; everyone knows coreference, they just don't know that they do. We trained a computer to know it too!

Coreference is a relationship between two or more expressions in a text when these expressions refer to the same person or thing. Coreference solving, the identification of sets of coreferring mentions in a text, is a well-studied problem in the field of natural language processing (NLP), the computational analysis of text. As an example, consider this short text: *John drove to Judy's house. He made her dinner.* which contains four noun phrases: *John*, *Judy*, *he* and *her*. A reader would intuitively connect *John* with *he* and *Judy* with *her* and surmise that *John* cooked dinner for *Judy*. By using linguistic terms we would say that the reader has solved the text's coreference and that the links the reader previously surmised were coreferring noun phrases.

In most cases the best coreference solvers are humans, but human labour has a high resource cost and would therefore be unfeasible for most tasks; it is often better to train a computer to do the work instead, even though the results are less impressive.

To train a coreference solver, one would need to gather a large collection of text containing manually annotated coreferences. The identified coreferences are then used to train a coreference solver by comparing coreferring and non-coreferring noun phrases. Some languages are fortunate with large amounts of training data, while some languages such as Swedish have very small or nonexistent data sets for this particular task. A good rule of thumb says that the minimum training size is in the vicinity of a million words. For Swedish, there exists only one data set with 20,000 words. Besides Swedish, many languages lack large training data sets.

In the absence of a large annotated data set, distant supervision offers a possible path forward. Distant supervision in the context of our Master's thesis means that we identify identical sentences in different languages, solve coreferences for one language, and try to map them to the other language. The initial solution or the transfer may be incorrect, but given sufficiently large texts the errors would hopefully be negligible.

The goal for our Master thesis is the creation of coreference solvers for the Swedish and German languages using this method.

Although the methods we describe have been used with some success in other languages, to the best of our knowledge, we are the first to create a coreference solver for Swedish using this technique.

We hope our results will pave the way for the creation of coreference solvers competitive with the current state of the art achieved by supervised training techniques.