Coreference Resolution for Swedish and German using Distant Supervision

Alexander Wallin
Lund University
Department of Computer Science
Lund, Sweden
alexander@wallindevelopment.se

Pierre Nugues
Lund University
Department of Computer Science
Lund, Sweden
Pierre.Nugues@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 paper, we describe a method for solving coreferences for Swedish and German using distant supervision that does not use manually annotated texts.

We generate a weakly labelled training set using parallel corpora, English-Swedish and English-German, where we solve the coreference for English using CoreNLP and transfer it to Swedish and German using word alignments. To carry this out, we identify mentions from dependency graphs in both target languages using hand-written rules. Finally, we evaluate the end-to-end results using the evaluation script from the CoNLL 2012 shared task for which we obtain a score of 34.98 for Swedish and 13.16 for German and, respectively, 46.73 and 36.98 using gold mentions.

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 referring words and phrases are generally called *mentions*. Coreference resolution is instrumental in many language processing applications such as information extraction, the construction of knowledge graphs, text summarizing, question answering, etc.

As most current high-performance coreference solvers use machine-learning techniques and supervised training (Clark and Manning, 2016), building solvers requires large amounts of texts, hand-annotated with coreference chains. Unfortunately, such corpora are expensive to produce and are far from being available for all the languages, including the Nordic languages.

In the case of Swedish, there seems to be only one available corpus annotated with coreferences: SUC-Core (Nilsson Björkenstam, 2013), which consists of 20,000 words and 2,758 coreferring mentions. In comparison, the CoNLL 2012 shared task (Pradhan et al., 2012) uses a training set of more than a million word and 155,560 coreferring mentions for the English language alone.

Although models trained on large corpora do not automatically result in better solver accuracies, the two orders of magnitude difference between the English CoNLL 2012 corpus and SUC-Core has certainly consequences on the model quality for English. Pradhan et al. (2012) posited that larger and more consistent corpora as well as a standardized evaluation scenario would be a way to improve the results in coreference resolution. The same should apply to Swedish. Unfortunately, annotating 1,000,000 words by hand requires seems to be out of reach for this language for now.

In this paper, we describe a distant supervision technique to train a coreference solver for Swedish and other languages lacking large annotated corpora. Instead of using SUC-Core to train a model, we used it for evaluation.

2 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 (Mintz et al., 2009; Yao et al., 2010). The primary difference between distant supervision and supervised learning lies in the annotation procedure of the training data;

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 46–55,
Gothenburg, Sweden, 23-24 May 2017.

supervised learning uses labelled data, often obtained through a manual annotation, whereas in the case of distant supervision, the annotation is automatically generated from another source than the training data itself.

Training data can be generated using various methods, such as simple heuristics or from the output of 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 advantage of distant supervision is that the training set does not need an initial annotation. Distant supervision covers a wide range of methods. In this paper, we used an annotation projection, where the output of a coreference resolver is transferred across a parallel corpus, from English to Swedish and English to German, and used as input for training a solver in the target language (Martins, 2015; Exner et al., 2015).

3 Previous Work

Parallel corpora have been used to transfer syntactic annotation. Hwa et al. (2005) is an example of this. In the case of coreference, Rahman and Ng (2012) used statistical machine translation to align words and sentences and transfer annotated data and other entities from one language to another. They collected 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.

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.

In this paper, we describe a new projection method, where we use a parallel corpus similarly to Hwa et al. (2005) and, where we follow the methods and metrics described by Rahman and Ng (2012). We also reused the maximum span heuristic in Martins (2015) and the pruning of documents according to the ratio between correct and incorrect entity alignments.

4 Approach

4.1 Overview

Our goal was to create a coreference solver for Swedish and German with no labelled data to train the model. Swedish has no corpora of sufficient size to train a general coreference solver, whereas German has a large labelled corpus in the form of Tüba D/Z (Henrich and Hinrichs, 2014). Although we could have trained a solver from the Tüba D/Z dataset, we applied the same projection methods to German to determine if our method would generalize beyond Swedish.

We generated weakly labelled data using a parallel corpus consisting of sentence-aligned text with a sentence mapping from English to Swedish and English to German. We annotated the English text using a coreference solver for English and we transferred the coreference chains to the target language by word alignment. We then used the transferred coreference chains to train coreference solvers for the target languages.

4.2 Processing Pipelines

We used three language-dependent processing pipelines:

- We applied Stanford's CoreNLP (Manning et al., 2014) to annotate the English part. We used the parts of speech, dependency graphs, and coreference chains;
- Mate Tools (Björkelund et al., 2010) for German;
- For Swedish, we used Stagger (Östling, 2013) for the parts of speech and MaltParser for the dependencies (Nivre et al., 2007).

4.3 Evaluation

As annotation and evaluation framework, we followed the CoNLL 2011 and 2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012). These tasks evaluated coreference resolution systems for three languages: English, Arabic, and Chinese. To score the systems, they defined a set of metrics as well as a script that serves as standard in the field.

We carried out the evaluation for both Swedish and German with this CoNLL script. For Swedish, we used SUC-Core (Nilsson Björkenstam, 2013) as a test set, while for German, we used the Tüba-D/Z corpus in the same manner as with SUC-Core (Henrich and Hinrichs, 2013; Henrich and Hinrichs, 2014).

5 Parallel Corpora

5.1 Europarl

As parallel corpora, we used the Europarl corpus (Koehn, 2005), consisting of protocols and articles

from the EU parliament gathered from 1996 in 21 language pairs.

Europarl is a large sentence-aligned unannotated corpus consisting of both text documents and web data in the XML format. Each language pair has alignment files to map the respective sentences in the different languages. We only used the text documents in this study and we removed unaligned sentences.

Koehn (2005) evaluated the Europarl corpus using the BLEU metric (Papineni et al., 2002). High BLEU scores are preferable as they often result in better word alignments (Yarowsky et al., 2001).

The BLEU 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, where 0 means no alignment and 100 means a perfect alignment. 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.

5.2 Word Alignment

To carry out the transfer of entity mentions, we aligned the sentences and the words of the parallel corpora, where English was the *source language* and Swedish and German, the *target languages*. Europarl aligns the documents and sentences using the Gale and Church algorithm. This introduces additional errors when aligning the words.

Instead, we used the precomputed word alignments from the open parallel corpus, OPUS, where improper word alignments are mitigated (Lee et al., 2010; Tiedemann, 2012). The word alignments in OPUS use the phrase-based grow-diag-final-and heuristic, which gave better results. Additionally, many of the challenges in aligning English to Swedish described by Ahrenberg (2010) appeared to be mitigated.

6 Bilingual Mention Alignment

From the word alignment, we carried out the mention transfer. We used a variation of the maximum span heuristic.

6.1 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.

Figure 1 shows two examples of good and bad projections from Yarowsky et al. (2001). 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. The solid lines correspond to word alignments while the dotted lines define the boundaries of their maximum span heuristic. Yarowsky et al. (2001) argue that even though individual word alignments are incorrect, a group of words corresponding to a noun phrase in the source language tends to be aligned with another group in the target language. 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.

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

The heuristic is nontrivial to evaluate and we primarily selected it for its simplicity, as well as its efficiency with coreference solvers for Spanish and Portuguese using distant supervision (Martins, 2015).

6.2 Maximum Span Optimal Mention

The maximum span heuristic uses no syntactic knowledge other than tokenization for the target language.

We implemented a variation of the maximum span heuristic which utilizes syntactic knowledge of the target language. We selected the largest mention bounded by each maximum span instead of the maximum span itself. As result, 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 and the problem of overlapping is replaced with pruning mentions with identical bounds.

6.3 Document Pruning

We removed some documents in the corpus from the generated data set according to two metrics: The document length and alignment agreement as in Martins (2015).

The goal was to create a training set with comparable size to the CoNLL task, i.e. a million

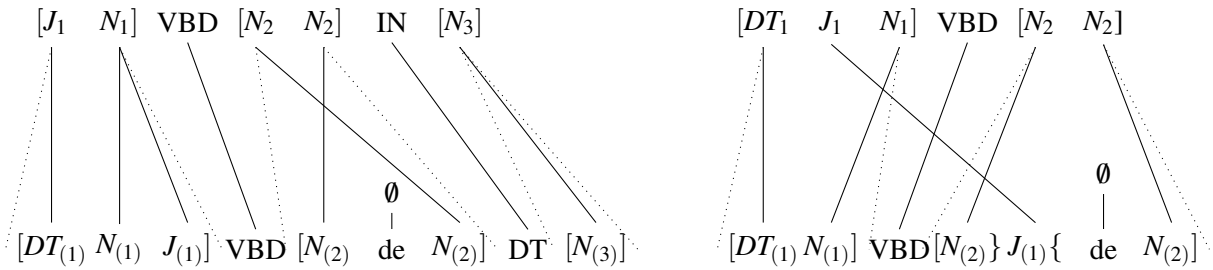

Figure 1: Left: Standard projection scenario according to Yarowsky et al. (2001); Right: Problematic projection scenario

words or more. To this effect, we aligned all the documents using the maximum span variant and we measured the alignment accuracy defined as the number of accepted alignments divided by the sum of all alignments.

We removed all the documents with lower than average alignment accuracy. Additionally, larger documents were removed until we could generate a total training set consisting of approximately a million words in total.

7 Evaluation

7.1 Metrics

There are multiple metrics to evaluate coreference resolution. We used the CoNLL 2012 score as it consists of a single value (Pradhan et al., 2012). This score is the mean of three other metrics: *MUC*6 (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_E$ (Luo, 2005). We also report the values we obtained with $CEAF_M$, and *BLANC* (Recasens and Hovy, 2011).

7.2 Test Sets

Swedish: SUC-Core. The SUC-Core corpus (Nilsson Björkenstam, 2013) consists of 20,000 words and tokens in 10 documents with 2,758 coreferring mentions. 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 coreference solver, but it is more than sufficient to evaluate solvers trained on some different source material. 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 have a CoNLL 2012 compatible file format. Additionally, we removed the singletons from the merged data files.

German: Tüba D/Z. The Tüba D/Z corpus (Henrich and Hinrichs, 2013; Henrich and Hinrichs, 2014) consists of 1,787,801 words and tokens organized in 3,644 files annotated with both part of speech and dependency graph information.

Although the corpus would be sufficient in size to train a coreference solver, we only used it for evaluation in this work. As with SUC-Core, we removed all the singletons. Due to time and memory constraints, we only used a subset of the Tüba D/Z corpus for evaluation.

7.3 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).

8 Experimental Setup

8.1 Selection of Training Sets

For Swedish, we restricted the training set to the shortest documents containing at least one coreference chain. After selection and pruning, this set consisted of 4,366,897 words and 183,207 sentences in 1,717 documents.

For German, we extracted a training set consisting of randomly selected documents containing at least one coreference chain. After selection and

pruning, the set consisted of 9,028,208 words and 342,852 sentences in 1,717 documents.

8.2 Mention Identification

Swedish. The mentions in SUC-Core correspond to noun phrases. We identified them automatically from the dependency graphs produced by Maltparser using a set of rules based on the mention headwords. Table 1 shows these rules that consist of a part of speech and an additional constraint. When a rule matches the part of speech of a word, we create the mention from its yield.

As SUC-Core does not explicitly define the mention bracketing rules, we had to further analyze this corpus to discern basic patterns and adjust the rules to better map the mention boundaries (Table 2).

German. The identification of noun phrases in German proved more complicated than in Swedish, especially due to the split antecedents linked by a coordinating conjunction. Consider the phrase *Anna and Paul*. *Anna*, *Paul*, as well as the whole phrase are mentions of entities. 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 the same kind of split antecedents only needs to check whether a conjunction has children that were noun phrases, whereas in German the same rule required more analysis.

Table 3 shows the rules for the identification of noun phrases in German, and Table 4, the post-processing rules.

9 Algorithms

9.1 Generating a Training Set

To solve coreference, we used a variation of the closest antecedent approach described in Soon et al. (2001). This approach models chains as a projected graph with mentions as vertices, 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.

The problem is framed as a binary classification problem, where the system only needs to decide whether a mention and its closest antecedent corefer (Soon et al., 2001). When generating a training set, the negative examples are more frequent than

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 1: Hand-written rules for noun phrase identification for Swedish based on SUC-Core. The rules are ordered by precedence from top to bottom

#	Description
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.

Table 2: Additional hand-written rules for post processing the identified noun phrases

POS	Additional rule	NP
NN	Depend. head has POS NN	No
	Otherwise	Yes
NE	Depend. 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 3: Hand-written rules for noun phrase identification for German based on Tüba-D/Z

#	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.

Table 4: Additional hand-written rules for post processing the identified noun phrases in German

the positive ones, which may skew the model. We limited the ratio at somewhere between 4 and 5 % and randomizing which negative samples become part of the final training set.

9.2 Machine-Learning Algorithms

We used the C4.5, random forest, and logistic regression algorithms from the Weka Toolkit and LibLinear to train the models (Witten and Frank, 2005; Hall et al., 2009; Fan et al., 2008).

9.3 Features

Swedish. As features, we used a subset Stamborg et al. (2012) and Soon et al. (2001). Table 5 shows the complete list.

German. The feature set for German is described in Table 5. The primary difference between German and Swedish is the addition of gender classified names. We used the lists of names and job titles from IMS Hotcoref DE (Rösiger and Kuhn, 2016) to train 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.

10 Results

10.1 Mention Alignment

Swedish. The Swedish EuroParl corpus consists of 8,445 documents. From these documents, we selected a subset consisting of 3,445 documents based on the size, where we preferred the smaller documents.

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

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

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

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

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

10.2 Mention Identification

Swedish. Using the rules described in Table 1, we identified 91.35% of the mentions in SUCCore. We could improve the results to 95.82% with the additional post processing rules described in Table 2.

Rule	Type	sv	de
Mentions are identical	Boolean	✓	✓
Mention head words are identical	Boolean	✓	✓
POS of anaphora head word is PN	Boolean	✓	
POS of antecedent head word is PN	Boolean	✓	
POS of anaphora head word is PM	Boolean	✓	
POS of antecedent head word is PM	Boolean	✓	
Anaphora head word has the morphological feat DT	Boolean	✓	
Antecedent head grammatical article	Enum	✓	
Anaphora head grammatical article	Enum	✓	
Antecedent grammatical number	Enum	✓	
Anaphora grammatical number	Enum	✓	
Checks if mention contains a word which is a male first name	Boolean		✓
Checks if mention contains a word which is a female first name	Boolean		✓
Checks if mention contains a word which is a job title	Boolean		✓
Checks if mention contains a word which is a male first name	Boolean		✓
Checks if mention contains a word which is a female first name	Boolean		✓
Checks if mention contains a word which is a job title	Boolean		✓
Number of intervening sentences between the two mentions. Max. 10.	Integer		✓
Grammatical gender of antecedent head word	Enum	✓	✓
Grammatical gender of anaphora head word	Enum	✓	✓
Anaphora head is subject	Enum		✓
Antecedent head is subject	Enum		✓
Anaphora has the morphological feature gen	Enum		✓
Antecedent has the morphological feature gen	Enum		✓
Anaphora has the morphological feature ind	Enum		✓
Antecedent has the morphological feature ind	Enum		✓
Anaphora has the morphological feature nom	Enum		✓
Antecedent has the morphological feature nom	Enum		✓
Anaphora has the morphological feature sg	Enum		✓
Antecedent has the morphological feature sg	Enum		✓

Table 5: The feature set used for Swedish (sv) and German (de)

German. Using the rules described in Table 3, we identified 65.90% of the mentions in Tüba D/Z. With the additional post processing rules described in Table 4, we reached a percentage of 82.08%.

10.3 Coreference Resolution

Table 6 shows the end-to-end results when using predicted mentions and Table 7 shows the results with the same pipeline with gold mentions. These latter results correspond to the upper bound figures we could obtain with this technique with a same

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 6: End-to-end results using predicted 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

Table 7: End-to-end results using gold mention boundaries

feature set.

11 Conclusion

In this paper, we have described end-to-end coreference solvers for Swedish and German that used no annotated data. We used feature sets limited to simple linguistic features easily extracted from the Swedish treebank and the German Tiger corpus, respectively. A large subset of the feature set of Stamborg et al. (2012) would very likely improve the results in this work.

The results in Tables 6 and 7 show that even though the dependency grammar based approach for identifying mentions yields a decent performance compared with CoNLL 2011 and 2012, a better identification and pruning procedure would probably significantly improve the results. This is manifest in German, where using the gold mentions results in a considerable increase of the scores: Table 7 shows a difference of more than 23 points compared with those in Table 6. This demonstrates that the large difference in scores between Swedish and German has its source in the methods used for mention identification rather than in the different feature sets or the correctness of the training set. This can be explained by the difficulty to predict mentions for German, possibly because of the differences in the dependency grammar format, as relatively few mentions were identified using their head elements.

The final results also show that classifiers based on J48 and random forests produced better scores than logistic regression.

Coreference resolution using weak labelled training data from distant supervision enabled us to create coreference solvers for Swedish and German, even though the mention alignment in the parallel corpora was far from perfect. It is difficult to compare results we obtained in this article with those presented in CoNLL, as the languages and test sets are different. Despite this, we observe that when using gold mention boundaries, we reach a $MELA_{CoNLL}$ score for Swedish that is comparable with results obtained for Arabic in the CoNLL-2012 shared task using the similar preconditions. We believe this shows the method we proposed is viable. Our results are, however, lower than those obtained for English and Chinese in the same task and could probably be improved with a better mention detection.

Acknowledgments

This research was supported by Vetenskapsrådet, the Swedish research council, under the *Det digitaliserade samhället* program.

References

Lars Ahrenberg. 2010. Alignment-based profiling of Europarl data in an English-Swedish parallel corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel

Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, et al. 2003. Tiger annotationsschema. Technical report, Universität des Saarlandes.

Amit Bagga and Breck Baldwin. 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.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 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.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August. Association for Computational Linguistics.

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

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the Stockholm Umeå corpus version 2.0. Technical report, Stockholm University.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Verena Henrich and Erhard Hinrichs. 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.

Verena Henrich and Erhard Hinrichs. 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)*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(03):311–325.

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

Jae-Hee Lee, Seung-Wook Lee, Gumwon Hong, Young-Sook Hwang, Sang-Bum Kim, and Hae-Chang Rim. 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.

Xiaoqiang Luo. 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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

André F. T. Martins. 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, July. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. 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 1003–1011. Association for Computational Linguistics.

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

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 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.

Altaf Rahman and Vincent Ng. 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.

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

Ina Rösiger and Jonas Kuhn. 2016. IMS HotCoref DE: A data-driven co-reference resolver for German. In *LREC*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Marcus Stamborg, Dennis Medved, Peter Exner, and Pierre Nugues. 2012. Using syntactic dependencies to solve coreferences. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 64–70. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 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.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 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.