Combining the output of two coreference resolution systems for two source languages to improve annotation projection

Yulia Grishina
Applied Computational Linguistics
FSP Cognitive Science
University of Potsdam
grishina@uni-potsdam.de

Abstract

Although parallel coreference corpora can to a high degree support the development of SMT systems, there are no large-scale parallel datasets available due to the complexity of the annotation task and the variability in annotation schemes. In this study, we exploit an annotation projection method to combine the output of two coreference resolution systems for two different source languages (English, German) in order to create an annotated corpus for a third language (Russian). We show that our technique is superior to projecting annotations from a single source language, and we provide an in-depth analysis of the projected annotations in order to assess the perspectives of our approach.

1 Introduction

Most of the recent work on exploiting coreference relations in Machine Translation focused on improving the translation of anaphoric pronouns (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012; Novák et al., 2015; Guillou and Webber, 2015), disregarding other types of coreference relations, one of the reasons being the lack of annotated parallel corpora as well as the variability in the annotated data. However, this could be alleviated by exploiting annotation projection across parallel corpora to create more linguistically annotated resources for new languages. More importantly, applying annotation projection using several source languages would support the creation of corpora less biased towards the peculiarities of a single source annotation scheme.

In our study, we aim at exploring the usability of annotation projection for the transfer of automatically produced coreference chains. In particular, our idea is that using several source annotations produced by different systems could improve the performance of the projection method. Our approach to the annotation projection builds upon the approach recently introduced by (Grishina and Stede, 2017), who experimented with projecting manually annotated coreference chains from two source languages to the target language. However, our goal is slightly different: We are interested in developing a fully automatic pipeline, which would support the automatic creation of parallel annotated corpora in new languages. Therefore, in contrast to (Grishina and Stede, 2017), we use automatic source annotations produced by two state-of-the-art coreference systems, and we combine the output of our projection method for two source languages (English and German) to obtain target annotations for a third language (Russian). Through performing the error analysis of the projected annotations, we investigate the most common projection errors and assess the benefits and drawbacks of our method.

The paper is organized as follows: Section 2 presents an overview of the related work and Section 3 describes the experimental setup. In Section 4, we give a detailed error analysis and discuss the results of our experiment. The conclusions and the avenues for future research are presented in Section 5.

2 Related work

Annotation projection is a method that allows for automatically transferring annotations from a well-studied (source) language to a low-resource (target) language in a parallel corpus in order to automatically obtain annotated data. It was first introduced in the work of (Yarowsky et al., 2001)

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 67–72,
Copenhagen, Denmark, September 8, 2017. ©2017 Association for Computational Linguistics.

	News		Stories		Total	
	EN	DE	EN	DE	EN	DE
Markables	486	621	429	414	915	1035
Chains	125	200	57	68	182	268

Table 1: Number of markables and coreference chains in the automatic annotations

	MUC			B^3			$CEAF_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
berkeley (EN)	49.5	41.4	45.0	38.9	27.8	32.1	45.9	40.4	42.9	44.7	36.5	40.0
CorZu (DE)	66.9	59.2	62.5	59.2	41.3	46.6	52.4	52.8	52.3	59.5	51.1	53.8

Table 2: Evaluation of the automatic source annotations

and then extensively exploited for different kinds of linguistic tasks, including coreference resolution. Specifically, several studies used annotation projection to acquire annotated data, such as (Postolache et al., 2006; Rahman and Ng, 2012; Martins, 2015; Grishina and Stede, 2015).

Thereafter, (Grishina and Stede, 2017) proposed a multi-source method for annotation projection: They used a manually annotated trilingual coreference corpus and two source languages (English-German, English-Russian) to transfer annotations to the target language (Russian and German, respectively). Although their approach showed promising results, it was based on transferring manually produced annotations, which are typically not available for other languages and, more importantly, can not be acquired large-scale due to the complexity of the annotation task.

3 Annotation projection experiment

In our experiment, we propose a fully automatic projection setup: First, we perform coreference resolution on the source language data and then we implement the single- and multi-source approaches to transfer the automatically produced annotations. We use the English-German-Russian unannotated corpus of (Grishina and Stede, 2017) as the basis for our experiment, which contains texts in two genres – newswire texts (229 sentences per language) and short stories (184 sentences per language). Furthermore, we use manual annotations present in the corpus as the gold standard for our evaluation. It should be noted that the manual annotations were performed according to the parallel coreference annotation guidelines of (Grishina and Stede, 2016) that are in general compatible with the annotation of the the OntoNotes corpus (Hovy et al., 2006) and are therefore suitable for our evaluation.

3.1 Coreference resolution on the source language data

Since the main goal of this experiment is to assess the quality of the projection of automatic annotations, first we need to automatically label the source language data. For the English side of the corpus, we chose the Berkeley Entity Resolution system (Durrett and Klein, 2014), which was trained on the English part of the OntoNotes corpus (Hovy et al., 2006) and achieves the average F1 of 61.71 on the OntoNotes dataset (Durrett and Klein, 2014). For the German side of the corpus, we use the state-of-the-art CorZu system (Tuggener, 2016) to obtain the source annotations, which achieves the average of 66.9 F1 on the German part of the SemEval 2010 dataset (Klenner and Tuggener, 2011).

Corpus statistics for the English and German datasets are presented in Table 1. Interestingly, CorZu was able to resolve slightly more markables and coreference chains in total than Berkeley (1035 vs. 915, 268 vs. 182 respectively). In particular, the numbers of found markables and chains in English and German diverge for the newswire texts, which further supports the claim that this part of the corpus contains more complex coreference relations than the short stories[1].

To estimate the quality of the automatically produced annotations, we evaluate the resulting dataset against the manually annotated English and German parts of the corpus (Table 2). As one can see from this table, CorZu and Berkeley do not perform equally good on our dataset: the average F1 of 53.8 for German as compared to the average F1 of 40.0 for English.

[1] As already stated in (Grishina and Stede, 2017), the newswire texts contain a larger percentage of complex noun phrases than the short stories.

	MUC			B^3			CEAF$_m$			Avg.		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
en-ru	51.7	32.6	39.8	40.6	19.6	26.0	45.7	31.3	37.0	46.0	27.8	34.3
de-ru	55.5	23.6	32.8	42.1	13.0	19.1	43.0	25.3	31.6	46.9	20.6	27.8
en,de-ru												
Setting 1	58.5	33.6	42.5	43.9	19.8	26.9	55.7	30.3	39.1	52.7	27.9	**36.2**
Setting 2	85.2	14.9	24.7	76.8	7.8	13.8	75.8	17.1	27.6	**79.3**	13.3	22.0
Setting 3	49.4	36.1	41.5	35.9	22.1	26.7	38.3	35.2	36.5	41.2	**31.1**	34.9

Table 3: Projection results from English and German into Russian

3.2 Annotation projection strategies

For our experiment, we implement a direct projection method for coreference as described in (Grishina and Stede, 2015). Our method works as follows: For each markable on the source side, we automatically select all the corresponding tokens on the target side aligned to it, and we then take the span between the first and the last word as the new target markable, which has the same coreference chain number as the source one. Since the corpus was already sentence- and word-aligned[2], we use the available alignments to transfer the annotations.

Thereafter, we re-implement the multi-source approach as described in (Grishina and Stede, 2017). In particular, they (a) looked at disjoint chains coming from different sources and (b) used the notion of chain overlap to measure the similarity between two coreference chains that contain some identical mentions[3]. In our experiment, we apply the following strategies from (Grishina and Stede, 2017):

1. Setting 1 (‘add’): disjoint chains from one source language are added to all the chains projected from the other source language;

2. Setting 2 (‘unify-intersect’): the intersection of mentions for overlapping chains is selected.

3. Setting 3 (‘unify-concatenate’): chains that overlap are treated as one chain starting from a certain percentage of overlap.

For both single- and multi-source approaches, we deliberately rely solely on word alignment information to project the annotations, in order to keep our approach easily transferable to other languages.

3.3 Results

To evaluate the projection results, we computed the standard coreference metrics – MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998) and CEAF (Luo, 2005) – and their average for each of the approaches (Table 3). As one can see from the table, the quality of projections from English to Russian outperforms the quality of projections from German to Russian by 6.5 points F1. Moreover, while Precision number are quite similar, projections from English exhibit higher Recall numbers.

As for the multi-source settings, we were able to achieve the highest F1 of 36.2 by combining disjoint chains (Setting 1), which is 1.9 point higher than the best single-source projection scores and constitutes almost 62% of the quality of the projection of gold standard annotations reported in (Grishina and Stede, 2017). We were able to achieve the highest Precision scores by intersecting the overlapping chains (Setting 2) and the highest Recall by concatenating them (Setting 3).

Finally, we evaluate the annotations coming from English and German against each other, in order to estimate their comparability and the percentage of overlap. Interestingly, we achieve 52.0 F1, with Precision being slightly higher than Recall (53.9 vs. 50.2), which shows the dissimilarity between the two projections.

4 Error analysis and discussion

Analyzing the errors coming from each of the source languages, we first looked at the percentage of transferred mentions (Table 4): Using our method we were able to automatically transfer 82.7% of all the source markable from English and only 57.6% of all the source markables from German; similarly, the percentage of the transferred chains is lower for German than for English. Interestingly, while CorZu performs better on the source dataset than Berkeley, the results for the annotations projected from a single source

[2] Sentence alignment was performed using HunAlign (Varga et al., 2007); word alignments were computed with GIZA++ (Och and Ney, 2003) on a parallel newswire corpus (Grishina and Stede, 2015).

[3] Computed as Dice coefficient.

are the opposite: Annotation projection from English to Russian performs better than from German to Russian. Our hypothesis is that the reason for the lower percentage of transferred annotations is the lower quality of word alignments for German-Russian as compared to English-Russian. Furthermore, since the original language of the texts was English, we presume that the German and Russian translations are closer to English and less similar to each other.

	English		German	
	#	%	#	%
Markables	757	82.7	596	57.6
Chains	182	100.0	227	84.7

Table 4: Transferred chains and markables

Since we do not have access to any gold alignment data, we estimate the quality of the word alignments by computing the number of unaligned tokens. Not surprisingly, we see a higher percentage of unaligned words for German-Russian than for English-Russian: 17.03% vs. 14.96% respectively, which supports our hypothesis regarding the difference in the alignment quality for the two pairs. Furthermore, we computed the distribution of unaligned words: The highest percentage of unaligned tokens disregarding punctuation marks are prepositions; pronouns constitute only 3% and 5% of all unaligned words for the alignments between English-Russian and German-Russian respectively. However, these numbers do not constitute more than 5% of the overall number of pronouns in the corpus.

Following the work of (Grishina and Stede, 2017), we analyse the projection accuracy for common nouns ('Nc'), named entities ('Np') and pronouns ('P') separately[4]: Table 5 shows the percentage of correctly projected markables of each type out of all the projected markables of this type. Our results conform to the results of (Grishina and Stede, 2017): For both languages, pronouns exhibit the highest projection quality, while common and proper nouns are projected slightly less accurately, which is probably due to the fact that pronouns typically consist of single tokens and are better aligned than multi-token common and proper names. Overall, for all the markables, the projection accuracy for English-Russian is around 10% better than projection accuracy for German-Russian.

	en-ru	de-ru
Nc	64.5	60.7
Np	70.5	66.6
P	83.6	76.5
All	65.1	55.6

Table 5: Projection accuracy for common nouns, proper nouns and pronouns (%)

Moreover, we compare the projected annotations across the two genres. Interestingly, the results for the two languages vary: While the average coreference scores for English-Russian are quite comparable (news: 34.2 F1, stories: 33.3 F1), the scores for German-Russian differ considerably (news: 30.8 F1, stories: 20.8 F1). We attribute this difference to the quality of the source annotations and the performance of the source coreference resolvers on different genres of texts.

5 Summary and outlook

In this study, we assessed the applicability of annotation projection in a scenario where we have access to two coreference resolvers in two source languages, the output of which is projected to a third language in a low-resource setting. Our results have shown that projection from two source languages is able to reach 62% of the quality of the projection of manual annotations and improves the projection scores by 1.9 F1. Moreover, using the output of two completely different coreference resolution systems, we observed the similar tendencies as while projecting gold standard annotations: Projection from English to Russian achieves higher scores than projection from German to Russian, and pronouns have the highest projection accuracy.

Another important finding is that better source annotations does not necessarily result in better projection scores, which can be explained by the different quality of word alignments for both language pairs. Having investigated this issue, we conclude that alignments between German and Russian contain more unaligned units than the alignments between English and Russian. Our next steps include examining the alignment quality in more detail, which would require establishing a gold standard set of alignments (for at least noun phrases).

[4] Using the automatic POS annotations already present in the corpus and provided by TreeTagger (Schmid, 2013).

Overall, we envision our future work in exploiting more than two source annotations as well as multiple coreference resolution systems for a single source language to improve the source coreference annotations. Specifically, we plan on applying our method on other language pairs and datasets, in order to explore its generalizabililty for a wider range of languages. Furthermore, we are interested in exploiting our approach as a first step to create coreference annotated corpora in new languages by providing automatically projected target coreference chains to human annotators for a subsequent validation.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 79–85.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *Transactions of the Association for Computational Linguistics*.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora, Beijing, China*. Association for Computational Linguistics, page 14.

Yulia Grishina and Manfred Stede. 2016. *Parallel coreference annotation guidelines*. University of Potsdam.

Yulia Grishina and Manfred Stede. 2017. Multi-source annotation projection of coreference chains: Assessing strategies and testing opportunities. In *Second Workshop on Coreference Resolution Beyond OntoNotes*. Association for Computational Linguistics, page 41.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1–10.

Liane Guillou and Bonnie Webber. 2015. Analysing ParCor and its translations by state-of-the-art SMT systems. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, page 24.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT (International Workshop on Spoken Language Translation); Paris, France; December 2nd and 3rd, 2010.*. pages 283–289.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human language technology conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, pages 57–60.

Manfred Klenner and Don Tuggener. 2011. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In *Proceedings of the international conference on Recent Advances in Natural Language Processing*. pages 178–185.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*. Association for Computational Linguistics, pages 252–261.

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

André Martins. 2015. Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. volume 1, pages 1427–1437.

Michal Novák, Dieke Oele, and Gertjan van Noord. 2015. Comparison of coreference resolvers for deep syntax translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, page 17.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of 5th international conference on Language Resources and Evaluation (LREC)*.

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*. Association for Computational Linguistics, pages 720–730.

Helmut Schmid. 2013. Probabilistic part-of speech tagging using decision trees. In *New methods in language processing*. Routledge, page 154.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zurich.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4* 292:247.

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*. Association for Computational Linguistics, pages 45–52.

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*. Association for Computational Linguistics, pages 1–8.