Knowledge-lean projection of coreference chains across languages

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

Manfred Stede
Applied Computational Linguistics
University of Potsdam
stede@uni-potsdam.de

Abstract

Common technologies for automatic coreference resolution require either a language-specific rule set or large collections of manually annotated data, which is typically limited to newswire texts in major languages. This makes it difficult to develop coreference resolvers for a large number of the so-called low-resourced languages. We apply a direct projection algorithm on a multi-genre and multilingual corpus (English, German, Russian) to automatically produce coreference annotations for two target languages without exploiting any linguistic knowledge of the languages. Our evaluation of the projected annotations shows promising results, and the error analysis reveals structural differences of referring expressions and coreference chains for the three languages, which can now be targeted with more linguistically-informed projection algorithms.

1 Introduction

Coreference resolution requires relatively expensive resources, usually in terms of manual annotation. To alleviate this problem for low-resourced languages, techniques of annotation projection can be applied. In this paper, we report on experiments with projecting nominal coreference chains across bilingual corpora. Our goal is to see how well a knowledge-lean projection algorithm works for two relatively similar languages (English-German) and for less similar languages (English-Russian). Furthermore, we are interested in differences incurred by the text genre and therefore use three different genres: argumentative newspaper articles, narratives, and medicine instruction leaflets.

Our general aim is to explore the limitations of a knowledge-lean approach to the problem, so that it is easy to generalize to other low-resourced languages. For the annotation of the corpus, we created common annotation guidelines that make few assumptions on the structural features of the target languages. We used the guidelines to annotate texts of the three genres in the three languages, and provide results on inter-annotator agreement (see Section 3). For projection, we use a procedure based on sentence and word alignment as calculated by a standard tool (GIZA++) that was trained on corpora of moderate size. Thus at this point we deliberately do not apply linguistic knowledge on the languages involved. The experiments and results are described in Section 4. We present a qualitative error analysis showing that a number of structural divergences are responsible for many of the problems; this suggests that limited syntactic knowledge can be helpful for improving performance in follow-up work. Section 5 compares our results to the most closely related earlier work, and Section 6 concludes.

2 Related work

A *projection* approach is used to automatically transfer different types of linguistic annotation from one language to another. The idea of mapping from well-studied languages to low-resourced languages was initially introduced in the work of Yarowsky et al. (2001), who studied the induction of PoS and NE taggers, NP chunkers and morphological analyzers for different languages using annotation projection. Thereafter, the technique has been used for a variety of

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 14–22,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics

tasks, including PoS tagging and syntactic parsing (Hwa et al., 2005; Ozdowska, 2006; Tiedemann, 2014), semantic role labelling (Padó and Lapata, 2005), sentiment analysis (Mihalcea et al., 2007), mention detection (Zitouni and Florian, 2008), or named-entity recognition (Ehrmann et al., 2011).

To our knoweldge, the first application to coreference is due to Harabagiu and Maiorano (2000), who experimented with manually projecting coreference chains from English to Romanian using a translated parallel corpus. They showed that a coreference resolver trained on a parallel corpus can achieve better results than one trained on monolingual data. Then, Postolache and colleagues (2006) used automatic word alignment to project coreference annotations for the same data. Their goal was to achieve high precision, and thus they discarded from projection those referring expressions (henceforth: REs) whose syntactic heads were not properly aligned. Their results indeed show high precision (over 95%), but considerably lower recall (around 70%). We will discuss their approach in relation to ours in Section 5.

Mitkov and Barbu (2002) performed anaphora resolution using projection on a parallel English-French corpus, which lead to an improvement in the success rate of roughly 4% for both English and French. (Sayeed et al., 2009) used cross-lingual projection to improve the detection of coreferent named entities with the help of English-Arabic translations, and they reported better results than a monolingual resolver could achieve. (Rahman and Ng, 2012) used translation-based projection to train a coreference resolver, and achieved around 90% of the average F-scores of a supervised resolver in experiments with Spanish and Italian using few resources (only a mention extractor) for the target languages.

3 Multilingual coreference corpus

3.1 The corpus

Our corpus consists of 38 parallel texts in English, German and Russian, belonging to three genres: newswire articles (7 texts per language), short stories (3 texts per language), and medicine instruction leaflets (4 per language, only English-German)[1]. This choice is motivated by (i) the common observation that narrative texts are easier to process for coreference, (ii) the fact that news text is important for many applications, and (iii) the consideration of medical leaflets representing a somewhat "exotic" genre that exhibits many differences to the other two.

Corpus statistics are shown in Table 1. The stories contain more REs than the newswire texts, and the coreference chains of the stories tend to be much longer.

3.2 Annotation

Usually, coreference annotation guidelines have been designed with one target language in mind. In contrast, our goal was to have common guidelines for the three languages, in order to (i) obtain uniform nominal coreference annotations in our corpus (supporting the projection task), and (ii) facilitate extension to further languages. Regarding English, our guidelines are of similar length and quite compatible with the scheme used for OntoNotes - the largest annotated coreference corpus for the English language (Hovy et al., 2006). One exception is that we handle only NPs and do not annotate verbs that are coreferent with NPs.

Our guidelines borrow many decisions from the (relatively language-neutral) Potsdam Coreference Scheme (PoCoS) (Krasavina and Chiarcos, 2007), and we also considered the recently developed guidelines for thr English-German parallel corpus *ParCor* (Guillou et al., 2014). But it considers only pairwise annotation of anaphoric pronouns and their antecedents, whereas we annotate all REs appearing in a coreference chain (i.e. that are mentioned in the text at least twice).

For the time being, our annotation is restricted to the referential *identity*; we thus exclude cases of 'bridging' (also called 'indirect anaphora') or near-identity. The following types of REs are considered as markables: full NPs, proper names, and pronouns (personal, demonstrative, relative, reflexive, and pronominal adverbs). As in OntoNotes, generic nouns can corefer with definite full NPs or pronouns, but not with other generic nouns. In case of English nominal premodifiers, we only annotate a nominal premodifier if it can refer to a named entity (the $[\text{US}]_1$ politicians) or is an independent noun in the Genitive form ($[\text{creditor's}]_1$ choice); in all other cases,

[1] Newswire is taken from the multilingual newswire agency Project Syndicate (www.project-syndicate.org). Stories are taken from an online collection of parallel texts for second language acquisition (http://www.lonweb.org). Medical texts are from the EMEA subcorpus of the OPUS collection of parallel corpora (Tiedemann, 2009).

	Newswire			Stories			Medicine leaflets		Total		
	En	De	Ru	En	De	Ru	En	De	En	De	Ru
Tokens	5903	6268	5763	2619	2642	2343	3386	3002	11908	11912	8106
Sentences	239	252	239	190	186	192	160	160	589	598	431
REs	558	589	606	470	497	479	322	309	1350	1395	1085
Chains	124	140	140	45	45	48	90	88	259	273	188

Table 1: Statistics for the experimental corpus

nominal premodifiers are not annotated as separate markables (e.g., [bank account]).

When annotators identify a markable, they also record its RE type from an attribute menu. The markable span includes the syntactic head of the NP and all its modifiers, except for dependent relative clauses (because relative pronouns are treated as separate markables). As a divergence from OntoNotes, they have a separate relation for appositions, whereas we only include them in the head NP markable. Technically, we used the MMAX-2 coreference annotation tool[2], and the corpus was tokenized and split into sentences using the Europarl preprocessing tools[3]. Table 2 shows a breakdown of NP types of our markables for the three genres.

	Newswire	Stories	Med. leaflets
Named Entities	39.3	27.5	48.0
Personal pronouns	15.9	51.4	8.2
Definite NP	30.1	16.1	16.9
Relative pronouns	9.9	1.1	14.4
Indefinite NP	4.7	3.5	12.3
Other	0.1	0.4	0.2

Table 2: Types of NPs in the three genres (%)

3.3 Agreement

The English-German corpus was annotated by two lightly-trained independent annotators - students of linguistics. (For Russian, we had only one annotator available, therefore the agreement study will be done later.) For markables, we computed the inter-annotator agreement using Cohen's kappa in two settings: binary overlap and proportional overlap. For binary overlap, we consider two markables as "agreed" if they overlap by at least one token; proportional overlap measures the extent to which annotators agree on the identification of spans (number of overlapping tokens). For the coreference annotation, we computed MUC scores with strict mention matching. The results for the newswire texts and stories are shown in Table 3. For the medical leaflets, the results are somewhat lower: $\kappa = 0.76$ with binary overlap and 0.67 with proportional overlap; the MUC score is 70%. For the NP type attribute, Cohen's kappa for the texts from all genres on average is $\kappa = 0.94$.

	English	German
Binary overlap κ	0.87	0.86
Proportional overlap κ	0.81	0.81
MUC F-score	77.28	73.91

Table 3: Inter-annotator agreement for news and stories

4 Experiment

4.1 Experimental setup

Automatic sentence and word alignment. We aligned the source and target parts of the corpus at the sentence level using the HunAlign sentence aligner (Varga et al., 2007) and its wrapper LF Aligner[4], which already includes alignment dictionaries for the required language pairs.

Word alignment was performed with GIZA++ (Och and Ney, 2003) using the standard settings. Before the alignment, all texts in the corpus were tokenized and lower-cased using the Europarl preprocessing tools. The word aligner was trained on a collection of bilingual newswire text from our source given above, preprocessed in the same way as descibed above. The training set consists of around 200 000 parallel sentences for English-German, and 170 000 for English-Russian.

We computed both bidirectional alignments and the intersection of source-target / target-source alignments. (Annotation projection is often done with intersective alignments, as they provide higher precision than bidirectional alignments.) For English-German, we evaluated our word alignment against a set of 1000 manually annotated parallel sentences made available by S. Padó[5]. For English-Russian, we are not aware of any similar gold alignments and thus did not

[2] http://mmax2.sourceforge.net

[3] http://www.statmt.org/europarl/

[4] http://aligner.sourceforge.net

[5] http://nlpado.de/ sebastian/data/srl_data.shtml

evaluate. Results are given in Table 4. Following (Padó, 2007), we evaluated only the resulting intersective alignments. We compared our results to those of (Padó, 2007) and (Spreyer, 2011), who used the English-German part of the Europarl dataset. Our results are somewhat lower, probably due to the much smaller training set.

	Bisentences	Prec.	Recall	F-m.
Padó (2007)	1 029 400	98.6	52.9	68.86
Spreyer (2011)	1 314 944	94.88	62.04	75.02
Our alignment	205 208	92.95	51.23	66.05

Table 4: Evaluation of the automatic word alignment

To simplify subsequent processing, we converted the corpus annotations into the CoNLL table format[6] using discoursegraphs converter (Neumann, 2015).

Extraction of REs and transfer of coreference chains. For each RE in the source language we extract the corresponding RE in the target language, together with its coreference set number. Following the approach of Postolache et al. (2006), for each word span representing an RE in the source language, we extract the corresponding set of aligned words in the target language. The resulting target RE is the span between the first and the last extracted word, and it belongs to the same set as the source RE. Table 5 shows the number of REs and coreference chains projected through word alignment (from English).

4.2 Evaluation

We evaluate both the quality of the identification of mentions and the extraction of coreference chains using the CoNLL scorer[7].

1. Evaluation of the identification of mentions.

   We compute the scores for the identification of mentions using the strict mention matching as in the CoNLL-2011 (Pradhan et al., 2011) and CONLL-2012 shared tasks (Pradhan et al., 2012), so that we score only those projected markable spans that are exactly the same as the gold ones. The values for English-German and English-Russian are given in Table 6 as *mentions*.

2. Evaluation of coreference chains

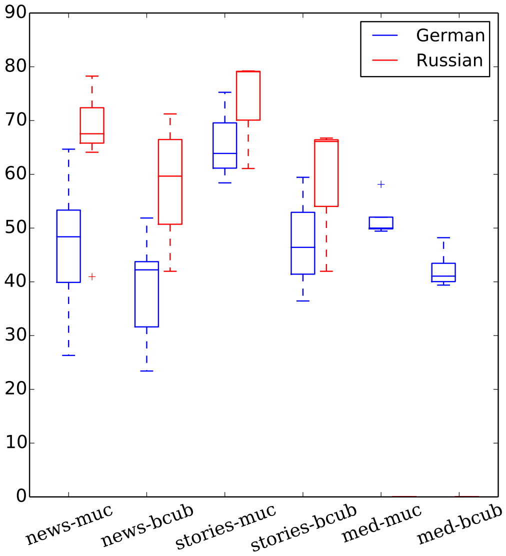

Figure 1: Comparison of English-German and English-Russian projections: boxplots of the macro-averaged F1 scores (MUC and B-cubed) for different genres

   We evaluate all the projected coreference chains against gold chains using the standard coreference evaluation metrics MUC (Vilain et al., 1995), CEAF (Luo, 2005) and B^3 (Bagga and Baldwin, 1998) to get complete performance characteristics. We also use strict matching as in the evaluation of the identification of mentions and evaluate the projected markables against all the markables of the gold standard. These scores depend on the identification of mentions evaluated in the previous step. We report the micro-averaged Precision, Recall and F-1 scores in Table 6. In addition, Figure 1 shows the distribution of macro-averaged F1-scores for two of the metrics (MUC and B^3) for both language pairs as boxplots.

3. Evaluation of coreference chains with minimal spans

   Finally, we evaluate using just minimal spans of the REs, i.e., syntactic heads. This indicates how well the REs can be projected, not punishing the algorithm for detecting only partially correct REs. We manually annotated syntactic heads of the gold and projected REs. Following the approach of Postolache et al. (2006), we select the leftmost

[6] http://conll.cemantix.org/2012/data.html
[7] http://conll.cemantix.org/2012/software.html

	Newswire		Stories		Medicine
	De	Ru	De	Ru	De
Transferred REs	465	493	329	357	214
Transferred coreference chains	122	122	44	44	82

Table 5: Number of REs and coreference chains transferred through bilingual projections

	Mentions			MUC			CEAF			B^3			Average (coref)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
de-News	61.5	48.6	54.3	55.9	43.2	48.7	58.6	46.7	51.9	45.8	34.2	39.1	53.4	41.4	46.6
de-Stories	82.0	54.5	65.5	81.9	51.6	63.3	81.7	53. 7	64.8	71.6	32.5	44.7	78.4	45.9	57.6
de-Medicine	61.2	44.7	51.7	66.2	42.7	51.9	59.1	43.3	50.0	53.43	35.16	42.41	59.6	40.4	48.1
de-News$_{min}$	89.9	71.2	79.4	87.3	66.2	75.3	85.5	67.5	75.5	80.4	58.1	67.5	84.4	63.9	72.8
de-Stories$_{min}$	95.4	62.2	75.3	94.4	58.5	72.2	95.1	61.2	74.5	90.9	40.2	55.7	93.5	53.3	67.5
de-Medicine$_{min}$	79.9	58.4	67.5	84.2	54.4	66.1	77.7	56.9	65.7	73.3	47.2	57.4	78.4	52.8	63.1
ru-News	79.3	64.5	71.2	76.3	60.7	67.6	76.3	62.0	68.4	69.0	52.2	59.4	73.9	58.3	65.1
ru-Stories	87.4	65.1	74.6	87.9	64.4	74.3	86.1	64.6	73.8	79.7	47.9	59.8	84.6	59.0	69.3
ru-News$_{min}$	90.9	72.6	80.7	89.6	69.8	78.5	87.3	69.7	77.5	83.7	61.4	70.9	86.9	67.0	75.6
ru-Stories$_{min}$	94.3	72.4	81.9	94.0	70.9	80.9	93.6	71.7	81.2	90.2	57.3	70.1	92.6	66.6	77.4

Table 6: Results for German and Russian: micro-averaged Precision, Recall, F1-score for different genres

noun, pronoun or numeral as head; otherwise, the RE is discarded. Results are given in Table 6 with the tag *'min'*.

4.3 Error Analysis

From a formal viewpoint, there are three categories of projection problems:

1. An RE is present in both source and target text, but it is not projected correctly, or not at all, on the grounds of mistakes in the word alignment phase.

2. An RE is present in the source text and correctly projected into the target text, but it does not show up in the gold standard, because the target language text does not have a corresponding RE *pair* (the target language does not reproduce the complete chain of the source).

3. An RE in the gold standard is not present in the target text and therefore can not be projected (the dual problem to (2): the source text does not have an RE pair that would correspond to one in the target text).

The number of errors caused by wrong word alignment (1) can be estimated on the basis of the alignment evaluation (Section 4.1), albeit only for the English-German language pair; due to the lack of resources, this is not possible for English-Russian.

Problems (2) and (3) are the more interesting ones for a qualitative error analysis. For this purpose, we visualized the projected files and the gold standard using the coreference module of the ICARUS corpus analysis platform (Gärtner et al., 2014). 50% of the data was randomly selected for the detailed analysis, and we determined the most frequent projection errors and categorized them into three different groups. Thereafter, we tried to verify our resulting hypotheses about variation in pronominal coreference in the three languages using a larger external corpus: InterCorp[8] (Čermák and Rosen, 2012) offers an online interface for searching parallel corpora in different languages and sub-corpora. We performed both monolingual and multilingual queries (e.g. querying one side of a parallel corpus vs. querying parallel data).

Further, we were interested in comparing our findings to available studies on multilingual nominal coreference in Contrastive Linguistics. However, the only work we found on this topic is a comparative study of nominal referring expressions for newswire texts in English and German (Kunz, 2010).

In our data, the problematic cases are those where the source language (SL) referring expression is missing or reformulated in the target text (TL), and therefore is not being projected. We identified three categories of errors caused by structural differences among the three languages:

[8] www.korpus.cz/intercorp.

Morphological differences.
These are cases of German contractions and compound nouns. For example, as in the case of *policy towards [minorities]*$_1$ and *[Minderheiten]politik*, the SL markable is not present in the TL as a separate unit, since we cannot split compound nouns and mark only a part. Also, cases like *zum Bahnhof* short for *zu dem Bahnhof* ('to the station') cause errors in the identification of spans, because we do not annotate prepositions as parts of markables on the English side. However, such cases are frequent in the German data, where, in general, the prepositions *an, bei, in, von, zu* can be contracted with subsequent determiners in written text. Our corpus study has shown that for the preposition *zu* ('to') the frequency of the contraction is 16 times higher than for the full form (InterCorp, measured in items per million (henceforth *i.p.m.*)).

Differences in NP syntax.
1: The use of articles. Some NPs are more frequently used with a definite article in German than in English, which resulted in the misidentification of spans. According to Kunz (2010), English allows the use of nouns with zero article more frequently than German. This is true for both singular and plural nouns. In our guidelines, nouns with zero article can only be linked to anaphoric pronouns (if any), but not between each other (like in OntoNotes). This resulted in mismatching chains: English NPs with zero article do not form chains and therefore cannot be projected, while the same NPs actually form a chain in German. For example:

(1) a. Lastly, the G-20 could also help drive momentum on *climate change*. <...> We also have to find a way to provide funding for adaptation and mitigation - to protect people from the impact of *climate change* and enable economies to grow while holding down pollution levels - while guarding against trade protection in the name of *climate change* mitigation.
b. Schließlich könnten die G-20 auch für neue Impulse im Bereich [des Klimawandels]$_1$ sorgen. Ebenso müssen wir einen Weg finden, finanzielle Mittel für die Anpassung an [den Klimawandel]$_1$ sowie dessen Eindämmung bereitzustellen - um die Menschen zu schützen und den Ökonomien Wachstum zu ermöglichen, aber den Grad der Umweltverschmutzung trotzdem in Grenzen zu halten. Außerdem gilt es, sich vor handelspolitischen Schutzmaßnahmen im Namen der Eindämmung [des Klimawandels]$_1$ zu hüten .

The query of InterCorp data has shown that German exhibits a higher number of NPs with definite article (57.928,55 i.p.m.) compared to English (31.405,22 i.p.m.). We also noticed that article use with named entities can vary in both languages (for example, the English *Hamas* corresponds to the German *die Hamas*). However, our corpus queries did not show any regularities yet; this issue requires a more detailed study regarding the types of named entities (which we assume to be the reason for the different use of articles). In the case of Russian, the absence of articles led to better results in the identification of REs, since in general, shorter spans increase the chance for a perfect alignment.

2: The use of reflexive pronouns. According to our annotation scheme, we annotated reflexive pronouns only when they are independent constituents (rather than verb particles), but we observe differences in the use of these pronouns for the three languages, so that in most cases these are non-parallel. These differences have to do with the form and distribution of reflexive pronouns. In English, we only have *-self* to express reflexivity, while in German and Russian a wider range of reflexives can be used. In German and Russian, it is possible to use more than one reflexive in a sentence to emphasize the action, which is not possible in English. As a result, there is less reflexives to be transferred from English to the target (German and Russian) sides of the corpus which led to errors in the projection.

3: Pre- and post-modification. In general, we noticed that German NPs allow more complicated premodification than English and Russian. According to Kunz (2010), English tends to postmodification, while German is less restrictive with premodification. These variations result in syntactical differences in markables and in non-parallelism.

Regarding the participial constructions, one of the complications is that in German, they occur only in pre-position, while in English and Russian they can be placed in both pre- and post-position. For example:

(2) a. Pakistan needs international help to bring hope to [*the young people*]$_1$ [*who*]$_1$ live there.
b. Pakistan braucht internationale Hilfe, um [*den dort lebenden jungen Menschen*]$_1$ Hoffnung zu bringen.

Non-equivalences in translation. The following cases of non-parallelism resulted in projection errors in our dataset; however, we could not find enough evidence to characterize them as systematic.

- Personal pronouns vs. indefinite pronouns.

  (3) a. [*It*]$_1$ was pursuing a two-pronged strategy.
  b. [*Man*] verfolgte eine Doppelstrategie. ('One followed a two-pronged strategy.')

  The German indefinite pronoun *man* is the target of the projected annotations, but it is not a markable according to our guidelines: it is non-referring and thus unable to participate in RE chains.

- Possessive NPs vs. adjectives. Some possessive NPs in the SL (for example, *the government of [India]*$_1$) can be expressed through adjectives in the TL (*die [indische] Regierung* or *indijskoe pravitel'stvo ([индийское] правительство)*) and therefore are no markables.

- Determiners vs. possessive pronouns. Personal pronouns in English can be translated as articles in German (for example, *[its]*$_1$ *broader goal = das weiter gefasste Ziel*), so that the source RE has no correspondent in the TL. For Russian, in this case a possessive form of a reflexive pronoun *svoj (свой)* can be used, or the possessive pronoun can be omitted.

- Relative clauses in one language can correspond to participle constructions or PPs in another. Examples:
  a. [*a fat lady*]$_1$ [*who*]$_1$ wore a fur around her neck
  b. [*eine dicke Dame mit einer Pelzstola*]$_1$ ('a fat lady with a a fur')

4.4 Comparing the genres

According to Table 6 and Figure 1, we see that newswire texts get the lowest scores, the reason most likely being the more complicated NPs. In setting 2 (evaluation of minimal spans), both newswire texts and stories obtain closer F1-scores, but the stories still have better precision scores. The medicine instruction leaflets in setting 2 have the worst results, and we observe lower improvement for precision between two settings compared to the newswire texts. This indicates that the quality of coreference resolution for medical texts depends to a higher degree on the coreference relations, than on the identification of mentions. In these texts, we frequently find borderline cases of non-/reference, when dieseases, parts of the body, etc. are being mentioned. Here, we will try to make the annotation guidelines more specific.

5 Discussion

The most closely related work is the approach of (Postolache et al., 2006), but some differences are noteworthy. In contrast to Postolache and colleagues, we do not focus on maximising precision; instead, our goal is to assess how well projection can work for all the annotations. In general, we use neither language-dependent software nor any additional linguistic information about the target language in the coreference projection and evaluation. Postolache et al., in contrast, applied a dedicated Romanian-English word aligner[9] (which achieves an F-score of 83.3% compared to our 66.05% of the language-independent GIZA++) and used special rules that rely upon the POS information and syntactic heads to produce their annotations, and then discarded the incorrectly projected ones (we used such rules only in the evaluation of the projected heads of REs). These rules reduced the number of gold and projected REs in the English-Romanian corpus considerably: from 3422 to 2491 (Postolache et al., 2006).

In our case, we use *all* REs to evaluate the spans of the projected annotations and the resulting coreference chains. Comparing our evaluation to Postolache's evaluation of all REs, we can see that our results yield a higher MUC precision for all of the genres (average 68.0 for English-German, 82.1 for English-Russian vs. 52.3 for English-Romanian), but a lower recall for both languages (45.8/62.6 vs. 82.04), which results in different F-measure (Postolache et al. obtained an average F1 of 63.9 compared to our F1 of 54.6 for German and 71.0 for Russian). This can be explained by the lower quality of our automatic English-German alignments compared to

[9] The COWAL word aligner is a lexical aligner which is adjusted only for Romanian-English and requires a corpus with morpho-syntactic annotations (Tufis et al., 2006).

the English-Romanian; the Russian REs were extracted slightly more accurately due to the structural differences in NPs. We also observed different scores for newswire texts, stories and medical leaflets, while Postolache et al. only used texts of one genre and in fact one author (different chapters of the same fiction book).

Keeping these different parameters in mind, in order to compare our results in a fair way, we evaluated the RE heads following the same rules to extract minimal spans of the projected REs, and evaluated them against manually annotated heads in the gold standard. In this setting, we obtained higher precision than in the previous setting, and in comparison to Postolache et al. (English-Romanian, avg. F1 = 80.5), our results are somewhat lower for English-German (avg. F1 = 74.1) and slightly better for English-Russian (avg. F1 = 81.3), which we attribute to the overall more difficult (and therefore more generalizable) projection scenario in our approach.

6 Conclusions

The goal of this study was to explore to what extent the coreference projection task can be tackled with a decidedly "light weight" approach. In contrast to earlier work, we used a well-known, standard word alignment tool trained on a corpus of moderate size. Furthermore, we deliberately worked with projecting English annotations to two relatively different languages, Russian and German, in order to study the limitations of the approach. In order to be as "generalizable" as possible (especially for other low-resourced languages), we work on the basis of common, relatively lean, annotation guidelines for coreference, which make few assumptions on the specifics of the languages considered here.

We compared our results quantitatively to the most closely related work and argued that they are competitive, in particular because our task setting is more target-language-neutral, we used three languages rather than two, and we worked on three different genres of text.

Our qualitative error analysis showed that problems are due to a set of structural differences of NPs in the three languages. Having completed this "light-weight" study, we will now move forward by introducing limited syntactic knowledge of the languages involved (NP chunking) and explore how much performance can be gained in that way. Still, our emphasis remains on devising procedures that are generalizable to other low-resourced languages, so we will do these extensions in small steps only.

Our annotation guidelines and other material will be made available via our website http://www.ling.uni-potsdam.de/acl-lab/.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *RANLP*, pages 118–124.

Markus Gärtner, Anders Björkelund, Gregor Thiele, Wolfgang Seeker, and Jonas Kuhn. 2014. Visualization, search, and error analysis for coreference annotations. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3191–3198. European Language Resources Association.

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the sixth conference on Applied natural language processing*, pages 142–149. Association for Computational Linguistics.

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*, pages 57–60. Association for Computational Linguistics.

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.

Olga Krasavina and Christian Chiarcos. 2007. PoCoS: Potsdam coreference scheme. In *Proceedings of the Linguistic Annotation Workshop*, pages 156–163. Association for Computational Linguistics.

Kerstin Anna Kunz. 2010. *Variation in English and German Nominal Coreference: A Study of Political Essays*, volume 21. Peter Lang.

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.

Rada Mihalcea, Carmen Banea, and Janyce M. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of ACL-2007.*

Ruslan Mitkov and Catalina Barbu. 2002. Using bilingual corpora to improve pronoun resolution. *Languages in contrast*, 4(2):201–211.

Arne Neumann. 2015. discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 309.

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

Sylwia Ozdowska. 2006. Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In *Proceedings of the International Workshop on Cross-Language Knowledge Induction*, pages 53–60. Association for Computational Linguistics.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 859–866. Association for Computational Linguistics.

Sebastian Padó. 2007. *Cross-lingual annotation projection models for role-semantic information*. Ph.D. thesis, German Research Center for Artificial Intelligence and Saarland University.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of LREC-2006.*

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.

Asad Sayeed, Tamer Elsayed, Nikesh Garera, David Alexander, Tan Xu, Douglas W. Oard, David Yarowsky, and Christine Piatko. 2009. Arabic cross-document coreference detection. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 357–360. Association for Computational Linguistics.

Kathrin Spreyer. 2011. *Does it have to be trees?: Data-driven dependency parsing with incomplete and noisy training data*. Ph.D. thesis, Universitätsbibliothek Potsdam.

Jörg Tiedemann. 2009. News from OPUS-a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proc. COLING.*

Dan Tufis, Radu Ion, Alexandru Ceausu, and Dan Stefanescu. 2006. Improved lexical alignment by combining multiple reified alignments. In *EACL.*

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 Linguistics 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*, pages 45–52. 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.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 600–609. Association for Computational Linguistics.