Cross-lingual Coreference Resolution of Pronouns

Michal Novák and Zdeněk Žabokrtský
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800
{mnovak,zabokrtsky}@ufal.mff.cuni.cz

Abstract

This work is, to our knowledge, a first attempt at a machine learning approach to cross-lingual coreference resolution, i.e. coreference resolution (CR) performed on a bitext. Focusing on CR of English pronouns, we leverage language differences and enrich the feature set of a standard monolingual CR system for English with features extracted from the Czech side of the bitext. Our work also includes a supervised pronoun aligner that outperforms a GIZA++ baseline in terms of both intrinsic evaluation and evaluation on CR. The final cross-lingual CR system has successfully outperformed both a monolingual CR and a cross-lingual projection system.

1 Introduction

Coreference resolution (CR) is a well-established task in the field of Natural Language Processing (NLP). The majority of papers published so far has focused on the monolingual CR, mostly experimenting on the English data. An important step towards multilingual CR was the CoNLL-2012 Shared Task in Modeling Multilingual Unrestricted Coreference in OntoNotes, where the participants were asked to build a CR system that could be applied on three typologically different languages contained in the OntoNotes corpus (Hovy et al., 2006): English, Chinese, and Arabic.

Same just as in other NLP tasks such as part-of-speech tagging or parsing, recent years have witnessed a rising interest in cross-lingual projection techniques, mostly aiming at under-resourced languages.

However, little attention is paid to leveraging cross-lingual information for CR in two resource-rich languages. This is probably due to lack of bilingual resources annotated with coreference since such techniques would require rich linguistic annotation on both sides of the bitext. Moreover, to solve this issue using a supervised learner, one needs the gold standard of coreference at least on the target side of the bitext. On the other hand, given such data, the typological differences in languages can be exploited to aid a CR system to perform better than if CR is performed independently for each language.

The motivation for solving this task is threefold. Firstly, even though Statistical Machine Translation (SMT) has been attracting interest of the community for years, most systems do not take information beyond the sentence boundary into account, leaving the issues of discourse coherence unresolved. Having a better-quality bitext with coreference resolved could drive research in discourse-aware SMT forward. Secondly, although inter-sentential relations are neglected in SMT, current phrase-based system unintentionally resolve some of the coreference links within the sentence, using just the power of phrases. This might be leveraged by using the SMT output instead of a human-translated output in a cross-lingual CR scenario. Finally, even monolingual CR may be improved by applying semi-supervised learning methods in a smart way on a large bilingual corpus with automatic rich annotations, such as CzEng 1.0 (Bojar et al., 2012).

Our work examines cross-lingual CR on the Czech-English language pair. We focus on CR of English pronouns, particularly the 3rd person *central pronouns*. Central pronouns is a term coined by Quirk (1985) embracing personal, possessive and reflexive pronouns. For the sake of simplicity, we will denote

3rd person central pronouns by the word *pronouns* in the following. We ignore noun phrase coreference for two reasons. First, there has been no data set available for the Czech-English language pair with noun phrase coreference annotated, yet. Second, the language differences between languages show more clearly on pronouns than on nouns, as pronouns tend to be more constrained by various grammar rules across different languages.

Czech and English are typologically distant languages, which is also reflected in different behavior of pronouns. A cross-lingual CR system could substantially benefit from the necessity of the anaphor and its antecedent to agree in gender. Czech uses grammatical genders which are more evenly distributed among nouns than the notional genders[1] used in English, where male and female gender[2] are solely allocated to living objects. However, benefiting from the pronoun's gender becomes problematic for personal pronouns in subject position which are usually dropped from the surface representation in Czech. If their governing verb is in the past tense, the correct gender can be reconstructed from its form. With the verb in present or future tense, the pronoun's gender remains hidden. Possessive pronouns are used to a greater extent in English than in Czech. Same as articles, they play the role of determiners whereas in Czech, the determination and possession must be understood from the context. A missing Czech counterpart of an English possessive pronoun may indicate its antecedent to be in the same sentence. Moreover, Czech uses reflexive possessive pronouns, whose antecedent is easier to detect than for non-reflexive pronouns. On the other hand, English reflexive pronouns, unlike the Czech, carry gender and number information the resolver can benefit from.

In this work, we make to our knowledge a first attempt to leverage the language differences using a machine learning approach to improve CR on bitexts. To achieve this goal, we create a supervised CR model, proposing two sets of cross-lingual features: projected features used for Czech CR and an indicator feature of a projected Czech coreference link obtained by a Czech CR system. Note that for the latter set (actually comprising only a single feature), the Czech CR system would require gold annotation of Czech coreference. We did not consider new features that would address specific Czech-English correspondences.

The fact that a Czech counterpart is missing for many English pronouns has a negative effect on traditional unsupervised alignment approaches. We address this issue by a supervised aligner of pronouns that incorporates the result of the traditional aligner as a feature and adds other features that help detect the true Czech counterparts of English pronouns.

The structure of this paper is as follows: After introducing related work in Section 2 and describing the data used in experiments in Section 3, we present the design of a supervised approach to improve English pronoun alignment in Section 4. Section 5 describes the cross-lingual CR system and the experiments conducted with it. Finally, we discuss the main observations made in the experiments in Section 6 and conclude the paper in Section 7.

2 Related work

The task of coreference resolution has been studied for a few decades, with supervised systems dominating the field. The most popular approaches have been thoroughly summarized by Ng (2010).

The system for English CR we use has been built for automatic coreference annotation in the Czech-English parallel treebank CzEng 1.0 (Bojar et al., 2012). It is an implementation of the so-called mention ranking model, first introduced by Denis and Baldridge (2007).

Parallel bilingual data is often exploited to solve well-known tasks such as part-of-speech tagging (Das and Petrov, 2011), named entity recognition (Kim et al., 2012), name tagging (Li et al., 2012), and semantic role labeling (Zhuang and Zong, 2010). Undoubtedly, this approach is most popular with parsing. Joint parsing of both the source and the target text along with searching for the best alignment between the trees has been approached in a more (Burkett et al., 2010) or less (Smith and Smith, 2004; Burkett and Klein, 2008) integrated approach. However, much closer to our work is the research on

[1] "Nouns are classified semantically according to their coreferential relations with personal, reflexive and wh-pronouns." (Quirk et al., 1985, p.314)

[2] Quirk (1985) uses these terms instead of terms masculine and feminine related to grammatical gender.

bilingually-informed parsing by Haulrich (2012), in which English trees are used to enrich the feature set for a Danish parser and vice-versa. Rosa et al. (2012) explored the same approach on the Czech-English language pair. Moreover, they adapted this technique to parse the output of an SMT system.

As for coreference resolution in a bilingual scenario, most works focus on coreference projection (de Souza and Orsan, 2011; Rahman and Ng, 2012; Ogrodniczuk, 2013). Research on cross-lingual CR has been inhibited by the lack of coreference-annotated parallel corpora. There are only few such corpora, for instance an English-Romanian corpus containing full hand-annotated coreference chains including noun phrase coreference (Postolache et al., 2006) and two corpora with pronoun coreference annotations – Prague Czech-English Dependency Treebank 1.0 (Hajič et al., 2012, PCEDT) and the recently published English-German corpus ParCor 1.0 (Guillou et al., 2014).

However, the only attempts at cross-lingual CR date back to the time before these corpora were released. Harabagiu and Maiorano (2000) designed a CR system for English-Romanian bitexts while Mitkov and Barbu (2003) focused on the English-French language pair. Both extended their rule-based monolingual CR systems to apply some high-precision rules from one language to enhance the result in the other language. They both reported an improvement of about 4% in precision compared to the monolingual systems.

As concerns a machine learning approach, in the work by Veselovská et al. (2012), PCEDT was employed in related tasks – to identifying types of the English personal pronoun *it* and Czech types of the unexpressed subject. The tasks have been addressed by the isolated monolingual systems as well as by taking advantage of the features from the other language.

3 Main source of the data

As mentioned in Section 2, Czech is one of a few languages for which a coreference-annotated parallel corpus has been built – The Prague Czech-English Dependency Treebank (Hajič et al., 2012, PCEDT).[3]

PCEDT is a manually annotated Czech-English parallel treebank comprising over 1.2 million words for each language in almost 50,000 sentence pairs. The English part contains the entire Penn Treebank–Wall Street Journal Section (Linguistic Data Consortium, 1999) transformed into dependency trees, whereas the Czech part comprises the translations of all the texts from the English part. The data from both parts are annotated on three layers of linguistic description following the Prague tectogrammatics theory (Sgall, 1967; Sgall et al., 1986) – the morphological layer (where each token from the sentence gets a lemma and a POS tag), the analytical layer (surface syntax in the form of a dependency tree, where each node corresponds to a token in the sentence) and the tectogrammatical layer. Tectogrammatical representation of a sentence is a dependency tree, where only content words have their own nodes; on the other hand, it contains additional nodes, e.g., for pronouns unexpressed on the surface. This is also the layer where the coreference relations are annotated. PCEDT includes annotation of pronoun coreference and the so-called grammatical coreference[4] for Czech as well as English.

For the purpose of this work, we ignore all annotations originally provided by PCEDT. Annotations on the tectogrammatical layer, which is in the center of this work's attention, are mostly manual there. But to truly simulate the real-world scenario when given just a pair of parallel texts, we need to replace them with ones carried out in a fully automatic manner. The only two exceptions, where we employ the gold annotations, are the relations we aim to model, i.e. coreference links and our own annotation of alignment for English personal pronouns (see Section 4.1).

3.1 Fully Automatic Annotation

We have conducted automatic linguistic analysis on both the English and the Czech part of PCEDT, transforming the individual sentences into multi-layer dependency tree structures based on the Prague tectogrammatics theory. The analysis was carried out within the Treex framework (Popel and Žabokrtský, 2010).

[3] `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`

[4] Its antecedent is imposed by the grammar of the language, e.g. coreference of relative pronouns.

Treex is a multi-purpose open-source framework for NLP applications development, which integrates a wide range of modules, such as tools for sentence splitting, tokenization, morphological analysis, part-of-speech tagging, shallow and deep syntax parsing, named entity recognition, anaphora resolution, among others.

Moreover, we performed an unsupervised word alignment on the complete PCEDT using the MGIZA++ tool (Gao and Vogel, 2008), which is a multi-threaded version of the popular GIZA++ (Och and Ney, 2000) that supports applying a saved model on a new sentence pair. We used a model trained on CzEng 1.0, which is about 300 times bigger in terms of the number of sentence pairs. The resulting alignment of the `intersection` and `grow-diag-final-and` types was subsequently projected onto the tectogrammatical layer. Furthermore, a simple heuristic was applied to find the English counterparts for reconstructed Czech personal pronouns. We denote this alignment as the *original* in the following.

4 Supervised alignment

The alignment described in the previous section is sufficiently accurate for content words, such as verbs, nouns, and adjectives. However, errors become more frequent as we move to pronouns. Some reasons for this have already been outlined in Section 1, i.e. dropped subject personal pronouns and omitted possessive pronouns in Czech. In addition, English uses a pleonastic variant of the pronoun *it*, which also has no correspondence in Czech. Personal pronouns function in a sentence as a replacement of nouns. Thus, it is no exception if a pronoun is translated into a noun. And finally, the translation may be reworded to such an extent that the pronoun would carry no valuable information, and it disappears. All these cases are difficult for GIZA++ to tackle.

The pronoun correspondence problem has been already faced concerning the alignment of the personal pronoun *it* by Novák et al. (2013). The authors tried to find the Czech counterpart of *it* by taking the node that is aligned to the parent of *it* on the Czech side and picking the argument of the aligned node that agrees on the semantic role with the particular *it*. This approach assumed that the unsupervised alignment of the parent, which is likely to be a content word, is of higher quality than the alignment of *it* itself. Furthermore, it relied on high-accuracy semantic role labeling, which could only be justified because the experiments were conducted on data manually annotated with semantic roles.

As we are working with fully automatic annotations (i.e., much less reliable) and a wider range of words to align, we cannot just copy this rule-based approach. However, we can take a more robust approach of supervised machine learning and transform Novák et al.'s rule to one of the features in our alignment model.

In Section 4.1, we describe the manual annotation of alignment, then introduce the supervised model in Section 4.2, using features described in Section 4.3. Finally, we show the evaluation results of the alignment model in Section 4.4.

4.1 Manual Annotation of the Data

Supervised learning requires that the training data are manually labeled with a target variable. For this purpose, we set aside the section 19 of PCEDT. In this data, all occurrences of English personal pronouns have been coupled with its Czech counterpart by one human annotator. If no suitable Czech expression was found, the annotator identified a possible cause of the missing counterpart. The causes were then categorized into three classes – pleonastic *it*, missing possessive pronoun and missing correspondence due to translation rewording. So far, we do not distinguish these classes in our models and treat them in the same manner.

We managed to align 471 occurrences of personal pronouns, which account for over 50% of all occurrences in the section. The overall statistics of how English personal pronouns are translated into Czech is shown in Table 1.

It shows that more than 55% of English personal pronouns are dropped from the surface representation of the Czech sentence, though still present in its deep structure. In contrast, English pleonastic pronouns are not present even there. An interesting observation is that more than half of English possessives are either translated as reflexive possessives or completely missing in the Czech sentence. All these

CS\EN	personal	possessive	reflexive	Total
personal unexpressed	147	1		148
personal	37	2		39
demonstrative	17	1		18
noun	15	6		21
possessive	3	78		81
reflexive possessive		68		68
reflexive	1	2	5	8
other	6	1	3	10
pleonastic	24			24
reword	12	4		16
no possessive		38		38
Total	262	201	8	471

Table 1: The statistics on the correspondence of English personal pronouns to their Czech counterparts. The last three Czech categories indicate the reason why there is no corresponding word in Czech for an English pronoun.

phenomena might in the end be a source of helpful information to the CR system.

4.2 Model

The nature of the task of aligning a given English pronoun to its Czech counterpart is to pick the best-fitting one from a bunch of candidates. The set of candidates consists of all tectogrammatical nodes in the aligned Czech sentence. To allow the system to select no correspondence for a pronoun, we add a special candidate representing the null alignment.

We represent the candidate ranking task as a discriminative log-linear model trained in a cost-sensitive, one-against-all strategy with label-dependent features (`csoaa-ldf`) provided by the Vowpal Wabbit[5] machine learning toolkit. The feature weights are optimized by running stochastic gradient descent in 40 passes over the training data.

4.3 Features

The feature set consists of the following types of features, which consider an English pronoun and a Czech candidate from the corresponding Czech tree:

- **Original alignment features:** presumably the most valuable set of features. It indicates if there is a link between the two nodes in the original alignment and if there is any between their parents.

- **Graph features:** we designed these features to somehow reflect the distance between the nodes. The pair of aligned tectogrammatical trees is treated as a bipartite graph and a shortest path between the nodes is found using a sequence of dependency edges and a single alignment link. We applied the Dijkstra algorithm to find the shortest path. We ensure that it only uses a single alignment link by setting large weights to alignments and small weights to dependency edges, i.e., 100 and 1, respectively. The features then comprise the length of the shortest path and the sequence of edge labels (parent, child, alignment).

- **Grammatical features:** these include lemmas, part-of-speech tags, reflexivity indicators, semantic role labels both for each of the nodes individually and as a concatenation of the two.

- **Combined features:** these features combine selected features from the types mentioned above. The concatenation of parents' alignment and semantic role correspondence mimics the rule Novák et al. (2013) used to get better Czech counterparts for English *it* (see Section 4). Furthermore, features combining lemmas with direct alignment or alignment through parents are included.

[5] `https://github.com/JohnLangford/vowpal_wabbit/wiki`

Method	Train				Test			
	A	P	R	F	A	P	R	F
ORIGINAL	–	–	–	–	73.04	75.55	82.40	78.83
SUPERVISED	88.37	90.18	90.34	90.26	84.50	88.52	86.40	87.45

Table 2: Evaluation results of English-to-Czech pronoun alignment. The quality is measured in terms of accuracy (A), precision (P), recall (R), and F1-score (F).

4.4 Experiments and Results

The small amount of manually annotated data led us to evaluate alignment models by 10-fold cross-validation, with the results on the train and test partitions averaged over all folds.

We measured the quality of produced the alignment links in terms of both accuracy and F1-score, i.e., as the harmonic mean of precision and recall. While accuracy positively scores also the cases when a node is correctly labeled as having no alignment, precision and recall neglect these cases at all, thus describing how good a method is in finding the correct counterpart for a node.

Table 2 shows the performance of the supervised model with the best combination of features and learning method parameters and compares it to the original alignment described in Section 3.1. It shows an improvement of about 9% absolute in terms of both accuracy and F-score.

5 Cross-lingual coreference resolver for English

In this section, we describe cross-lingual coreference resolution. The CR system we use definitely does not aim to compete with current state-of-the-art systems. However, for the purpose of research on cross-lingual CR, it can be employed as a reasonable baseline.

In Section 5.1, we describe the supervised CR model trained and tested on the data described in Section 5.2. We elaborate more on the design of English and aligned features in Section 5.3 and Section 5.4, respectively. Finally, several variants of the CR system are evaluated and compared in Section 5.5.

5.1 Coreference model

Our resolver employs a supervised model denoted as *mention ranker* by Ng (2010). Its advantage lies in judging all antecedent candidates simultaneously, and then picking the candidate with the highest score as the predicted antecedent. However, it is unable to exploit features that describe already formed clusters of mentions belonging to the same entity. A typical issue related to ranking models is how to deal with non-anaphoric mentions. We use the approach introduced by Rahman and Ng (2009) – adding a special candidate that indicates no anaphor.

Since this work focuses only on the so-called pronoun resolution, all the anaphor candidates are English 3rd person central pronouns, i.e. personal, possessive and reflexive pronouns.

For every anaphor, we collect in the set of its antecedent candidates all semantic nouns[6] from the previous sentence and the part of the current sentence prior to the anaphor.

CR can be treated as a ranking task, so we represent it in the same way as we handled alignment in Section 3.1 – as a discriminative log-linear model trained in the `csoaa-ldf` strategy by the Vowpal Wabbit tool. The feature weights are optimized by running stochastic gradient descent in 20-80 passes (the number differs across the experiments) over the training data.

5.2 Data

Models for coreference were trained on data extracted from sections 00–18 of the automatically analyzed PCEDT (as described in Section 3). Sections 20–21 have been employed as development testing data and Sections 22–24 as evaluation testing data. The development set has been used to select the best configuration, which was subsequently tested on the evaluation set. The training, development, and evaluation set consist of 19,294, 1,988 and 2,591 instances with 86%, 67%, and 73% anaphoric instances, respectively.

[6] Semantic nouns are all nouns as well as pronouns acting as a noun.

5.3 English Features

A wide range of features used by us had already been proven to be beneficial for the task of CR in multiple prior works. The majority of the features presented here have already been used in the CR system for Czech (Nguy et al., 2009); we keep just the language-independent. Furthermore, several grammatical and positional features proposed by Charniak and Elsner (2009) have been added. Finally, the feature set has been enriched with the information on named entities and WordNet[7] classes. All the features disregard dependent members of a mention, describing just the head of the mention. They can be divided into several categories:

- **Distance features:** number of sentences, clauses, and words between the anaphor and the antecedent candidate; the order of the candidate,

- **Grammatical features:** morphological number and gender of both the anaphor and the antecedent candidate, agreement in gender and number; part-of-speech tag,

- **Function features:** they exploit dependency labels on the analytical layer and semantic roles on the tectogrammatical layer; they also include an indicator of whether the mention plays a role of an argument or an adjunct in the governing phrase,

- **Parent features:** the features of both nodes' parents, e.g. their lemmas or semantic roles, are compared; an indicator of whether a mention is in coordination,

- **Semantic features:** WordNet classes the head word is assigned to,

- **Named entity features:** the named entity category and subcategory returned by Stanford named entity recognizer.[8] This includes also the indicator of whether the mention is a name of a person,

- **Charniak features:** anaphor type (pronoun in subject position, in object position, possessive pronoun, reflexive pronoun, other); antecedent type (noun, pronoun, other); antecedent syntactic type (subject, object, prepositional phrase, other).

We denote this feature set as EN in all our experiments.

5.4 Alignment features

The features from the Czech nodes aligned to the given English anaphor and antecedent candidate are obtained by moving to the corresponding Czech nodes and extracting the features as though we are trying to resolve a Czech coreference link. As outlined in Section 1, we designed two sets of features: CS and CS-COREF.

The CS set consists of features introduced by Nguy et. al (2009). Most of them, namely the categories of distance, function, and parent features, are extracted in the same manner as the English ones in the previous section. Grammatical features also contain the full positional morphological tag as designed by Hajič (2004). Semantic features employ a different knowledge base, replacing WordNet by the Czech portion of EuroWordNet (Vossen, 1998). In addition to the features more or less shared with the English side, the Czech feature set includes a probability estimate of the antecedent candidate co-occurring with its governing verb. This statistics has been collected on Czech National Corpus (CNC, 2005).

The CS-COREF set consists of a single binary feature indicating if there is a coreference relation between the nodes predicted by the Czech CR system (Nguy et al., 2009), or not.

[7] `http://wordnet.princeton.edu`

[8] `http://nlp.stanford.edu/software/CRF-NER.shtml`

5.5 Experiments and Results

The different feature sets proposed in the previous sections suggest an obvious set of experiments. The system trained only on the monolingual EN features is put as a baseline.

The rest our experimental setups use alignment features, forming three combinations with EN features: EN + CS, EN + CS-COREF, and EN + CS + CS-COREF. Moreover, these three experiments can be run on the data provided either with the original or supervised alignment, which serves as extrinsic evaluation of alignment approaches. This allows us to confirm or deny the hypothesis that the alignment plays a significant role in cross-lingual CR (see Section 4).

For comparison, we also evaluated the system that simply projects coreference links obtained by the Czech CR system to English.

The performance of a CR system is usually measured by scores that treat CR as a clustering problem, e.g., MUC, B^3, CEAF. As this work focuses merely on a subset of coreference expressions – pronouns – and we only compare different feature sets trained in the same framework, we resorted to the simplest metrics with a sufficient expression power. For each English pronoun we test if its predicted antecedent hits any of the true antecedents within the window of the current and the previous sentence. Given this indicator we calculate precision, recall, and F1-score, which takes into account only the nodes for which a relation with another node exists – referential pronouns in this case (similarly to the alignment evaluation in Section 4.4). Likewise, in order to assess quality of detecting non-referential pronouns, accuracy is computed as well.

The final results are shown in Table 3. The overall higher numbers on the evaluation set than on the development set probably result from a different proportion of non-anaphoric pronouns (see Section 5.2). The smaller difference in F1-score than in accuracy also supports this explanation.

The coreference projection scores a great deal below the baseline, which suggests that this approach is worth using only if manual annotation for at least a small amount of target language data (English in our case) is extremely expensive.

As for the cross-lingual CR on the original alignment, all three feature set combinations have beaten the baseline. The EN + CS-COREF system confirmed the added value of the CS-COREF feature, which, unlike the CS feature set, conveys latent information on true Czech coreference links. Even the combination of all features performs worse than CS-COREF alone.

Moving to the experiments with supervised alignment, we can see the findings from Section 4.4 confirmed also in the extrinsic evaluation. All three systems outperform not only the baseline, but also all the systems working on the original alignment. Moreover, both accuracy and F1-score order the three feature combinations in the expected way, where the overall winner improves over the baseline in more than 1% absolute. This improvement is significant[9] at p-level $p \leq 0.1$ but not at p-level $p \leq 0.05$.

6 Discussion

Using information from Czech parallel texts in English CR led to an improvement in terms of automatic measures. To see what the main aspects in which the Czech text positively impacts the CR performance are, we compared the output of the system trained only on the EN features with systems working on the EN + CS and EN + CS-COREF feature sets. We used the results of the experiments run on the development set with supervised alignment for this comparison.

Out of 1988 coreference instances in the development set, the EN + CS system improved the output in 49 cases, while it worsened the output in 23 cases. The rest remained unchanged. Likewise, the EN + CS-COREF system scored better than the EN one in 63 instances, while it failed in 39 instances.

The inspection of 10% instances for which the systems differed revealed that the cases when the cross-lingual system scored better than the monolingual concur with the language differences described in Section 1. We found that in these cases, the pronoun is often a pleonastic *it* or a possessive pronoun with a Czech reflexive possessive counterpart. Finally, we noticed improvements in cases where the Czech antecedent is easier to determine due to agreement in gender and number.

[9] Significance has been calculated by bootstrap resampling using 100,000 samples.

Setup	Train				Dev				Eval			
	A	P	R	F	A	P	R	F	A	P	R	F
EN	79.13	80.12	86.00	82.96	60.97	60.28	79.14	68.43	63.72	63.28	78.78	70.19
Original alignment												
CS-COREF projection	28.64	49.57	21.75	30.23	36.55	41.98	24.66	31.07	33.33	42.38	21.58	28.60
EN + CS-COREF	78.31	79.27	85.25	82.15	61.77	61.07	80.45	69.44	64.30	63.74	79.62	70.80
EN + CS	83.32	84.05	89.97	86.91	61.97	61.15	80.23	69.40	64.07	63.72	78.62	70.39
EN + CS + CS-COREF	80.75	81.52	87.61	84.46	62.27	61.33	80.96	69.79	64.03	63.59	79.57	70.69
Supervised alignment												
CS-COREF projection	30.74	49.91	24.87	33.20	36.60	41.38	27.61	33.12	33.60	41.85	23.98	30.49
EN + CS	83.19	83.98	89.73	86.76	62.27	61.42	80.60	69.72	64.53	64.13	79.09	70.83
EN + CS-COREF	79.27	80.20	85.89	82.95	62.17	61.27	81.11	69.81	64.65	64.11	79.67	71.05
EN + CS + CS-COREF	81.99	82.78	88.53	85.56	62.68	61.59	81.62	70.20	64.69	64.38	79.67	71.22

Table 3: Evaluation results of monolingual CR, CR via projection, and cross-lingual CR system trained and tested on the data with both the original and supervised alignment. Performance is measured in terms of accuracy (A), precision (P), recall (R) and F1-score (F).

We did not encounter an example of improvement for an English possessive pronoun having no Czech counterpart. We might have inspected too little data for it to appear. However, these cases may get covered after the features combining English and Czech features will be introduced.

7 Conclusion

This work introduced a largely unexplored task in the field of CR – cross-lingual CR. Given a Czech-English bitext, we sought to improve the performance of an English pronoun CR system by enriching the feature set with features from the aligned Czech text. Consistent improvements over the monolingual system confirmed that cross-language differences in pronoun behavior are big enough to affect the result. Furthermore, we have found that the quality of alignment is vital for this task.

In future work, we plan to apply this approach on a much larger parallel corpus and employ semi-supervised techniques to improve cross-lingual as well as monolingual CR. Moreover, human translation in the bitext can be replaced with the output of SMT system to see if we can produce valuable features for CR from the machine-translated source text.

Acknowledgments

This work has been supported by the EU FP7 project Khresmoi (contract no. 257528). This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). We would like to thank our colleague Ondřej Dušek for language correction and three anonymous reviewers for their remarks and suggestions.

References

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. The joy of parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey. European Language Resources Association.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

CNC. 2005. Czech national corpus – SYN2005. Prague, Czech Republic. Institute of the Czech National Corpus.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Stroudsburg, PA, USA. Association for Computational Linguistics.

José Guilherme Camargo de Souza and Constantin Orsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, Berlin, Heidelberg. Springer-Verlag.

Pascal Denis and Jason Baldridge. 2007. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jrg Tiedemann, and Bonnie Webber. 2014. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, Reykjavik, Iceland. European Language Resources Association.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association.

Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Nakladatelství Karolinum.

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Wittorff Haulrich. 2012. *Data-driven bitext dependency parsing and alignment*. Ph.D. thesis, Copenhagen Business School.

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*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from Wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, New York, NY, USA. ACM.

Linguistic Data Consortium. 1999. Penn Treebank 3. LDC99T42.

Ruslan Mitkov and Catalina Barbu. 2003. Using bilingual corpora to improve pronoun resolution. *Languages in contrast*, 4(2).

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. 2009. Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK. The Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2013. Translation of "it" in a deep syntax framework. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Workshop on Discourse in Machine Translation*, Sofija, Bulgaria. Omnipress, Inc.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maciej Ogrodniczuk. 2013. Translation- and projection-based unsupervised coreference resolution for Polish. In *Language Processing and Intelligent Information Systems*, number 7912, Berlin / Heidelberg. Springer.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233, Berlin / Heidelberg. Springer.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. European Language Resources Association.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, Stroudsburg, PA, USA. 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*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rudolf Rosa, Ondřej Dušek, David Mareček, and Martin Popel. 2012. Using parallel features in parsing of machine-translated sentences for correction of grammatical errors. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht.

Petr Sgall. 1967. *Generativní popis jazyka a česká deklinace*. Academia, Prague, Czech Republic.

David A. Smith and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kateřina Veselovská, Giang Linh Nguy, and Michal Novák. 2012. Using Czech-English parallel corpora in automatic identification of “it”. In *The Fifth Workshop on Building and Using Comparable Corpora*, İstanbul, Turkey. European Language Resources Association.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Tao Zhuang and Chengqing Zong. 2010. Joint inference for bilingual semantic role labeling. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.