J. Hlaváčová (Ed.): ITAT 2017 Proceedings, pp. 193–200
CEUR Workshop Proceedings Vol. 1885, ISSN 1613-0073, © 2017 M. Novák

ITAT

Coreference Resolution System Not Only for Czech

Michal Novák

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, CZ-11800 Prague 1
mnovak@ufal.mff.cuni.cz

Abstract: The paper introduces Treex CR, a coreference resolution (CR) system not only for Czech. As its name suggests, it has been implemented as an integral part of the Treex NLP framework. The main feature that distinguishes it from other CR systems is that it operates on the tectogrammatical layer, a representation of deep syntax. This feature allows for natural handling of elided expressions, e.g. unexpressed subjects in Czech as well as generally ignored English anaphoric expression – relative pronouns and zeros. The system implements a sequence of mention ranking models specialized at particular types of coreferential expressions (relative, reflexive, personal pronouns etc.). It takes advantage of rich feature set extracted from the data linguistically preprocessed with Treex. We evaluated Treex CR on Czech and English datasets and compared it with other systems as well as with modules used in Treex so far.

1 Introduction

Coreference Resolution (CR) is a task of discovering coreference relations in a text. Coreference connects *mentions* of the same real-world entity. Knowing coreference relations may help in understanding the text better, and thus it can be used in various natural language processing applications including question answering, text summarization, and machine translation.

Most of the works on CR have focused on English. In English, a mention almost always corresponds to a chunk of actual text, i.e. it is expressed on the surface. But Czech, for instance, is a different story. Czech is a typical example of pro-drop languages. In other words, a pronoun in the subject position is usually dropped as it is in the following example: *"Honza miluje Márii. Taky <ZERO-on> miluje pivo." ("John loves Mary. He also loves beer.")* If we ignored Czech *subject zeros*, we would not be able to extract a lot of information encoded in the text.

But subject zeros are not the only coreferential expression that may be dropped from the surface. Indeed, such zero mentions may appear even in the language where one would not expect them. For instance, the following English sentence does not express the relative pronoun: *"John wants the beer <ZERO-that> Mary drinks."*

This paper presents the *Treex Coreference Resolver* (*Treex CR*).[1] It has been primarily designed with focus on resolution in Czech texts. Therefore, Treex CR naturally supports coreference resolution of zero mentions.

The platform that ensures this and that our system operates on is called *tectogrammatical layer*, a deep syntax representation of the text. It has been proposed in the theory of Prague tectogrammatics [32]. The tectogrammatical layer represents a sentence as a dependency tree, whose nodes are formed by content words only. All the function and auxiliary words are hidden in a corresponding content node. On the other hand, the tectogrammatical tree can represent a content word that is unexpressed on the surface as a full-fledged node.

T-layer is also the place where coreference is represented. A generally used style of representing coreference is by co-indexing continuous chunks of surface text. Tectogrammatics adopts a different style. A coreference link always connects two tectogrammatical nodes that represent mentions' heads. Unlike the surface style, tectogrammatics does not specify a span of the mention, though. Such representation should be easier for a resolver to handle as the errors introduced by wrong identification of mention boundaries are eliminated. On the other hand, for some mentions it may be unclear what its head is.[2]

At this point, let us introduce the linguistic terminology that we use in the rest of the paper. Multiple coreferential mentions form a chain. Splitting the chain into pairs of mentions, we can adopt the terminology used for a related phenomena – *anaphoric relations*. The anaphoric relation connects the mention which depends upon another mention used in the earlier context.[3] The later mention is denoted as the *anaphor* while the earlier mention is called the *antecedent*.

This work is motivated by cross-lingual studies of coreferential relations. We thus concentrate mostly on pronouns and zeros, which behave differently in distant languages, such as Czech and English.[4] Coreference of nominal groups is not in the scope of this work because it is less interesting from this perspective.

However, Treex CR is still supposed to be a standard coreference resolver. We thus compare its performance with three coreference resolvers from the Stanford Core

[1] It is freely available at `https://github.com/ufal/treex` as the module `Treex::Scen::Coref` in the Treex framework.

[2] As we demonstrate in Section 5.

[3] As opposed to *cataphoric relations*, where the dependence is oriented to the future context.

[4] A thorough analysis of correspondences between Czech and English coreferential expressions has been conducted in [26].

NLP toolkit, which are the current and former state-of-the-art systems for English. Since we evaluate all the systems on two datasets using the measure that may focus on specific anaphor types, this work also offers a non-traditional comparison of established systems for English.

2 Related Work

Coreference resolution has experienced evolution typical for most of the problems in natural language processing. Starting with rule-based approaches (summarized in [20]), the period of supervised (summarized in [23]) and unsupervised learning methods (e.g. [6] and [15]) followed. This period has been particularly colorful, having defined three standard models for CR and introduced multiple adjustments of system design. For instance, our Treex CR system implements some of them: mention-ranking model [10], joint anaphoricity detection and antecedent selection, and specialized models [11]. A recent tsunami of deep neural network appears to be a small wave in the field of research on coreference. Neural Stanford system [8] set a new state of the art, yet the change of direction has not been as rapid and massive as for the other, more popular, topics, e.g. machine translation.

The evolution of CR for Czech proceeded in a similar way. It started during the annotation work on Prague Dependency Treebank 2.0 [16, PDT 2.0] and a set of deterministic filters for personal pronouns proposed by [17], followed by a rule-based system for all coreferential relations annotated in PDT 2.0 [24]. Release of the first coreference-annotated treebank opened the door for supervised methods. A supervised resolver for personal pronouns and subject zeros [25] is the biggest inspiration for the present work. We use a similar architecture implementing multiple mention-ranking models [10] specialized on individual anaphor types [11]. Unlike [25], we use a richer feature set and extend the resolver also to other anaphor types.

Moreover, we rectify a fundamental shortcoming of all these coreference resolvers for Czech – the experiments with them were conducted on the manual annotation of tectogrammatical layer. In this way, the systems could take advantage of gold syntax or disambiguated genders and numbers. While the rule-based system [24] reports around 99% F-score on relative pronouns, fair evaluation of a similar method but run on automatic tectogrammatical annotation reports only 57% F-score (see Table 2). If the system uses linguistically pre-processed data, the preprocessing must always be performed automatically.

3 System Architecture

Treex Coreference Resolver has been developed as an integral part of the Treex framework for natural language processing [29]. Treex CR is a unified solution for finding coreferential relations on the t-layer. For that reason, it requires the input texts to be automatically pre-processed up to this level of linguistic annotation. The system is based on machine learning, thus making all the components fully trainable if the appropriate training data is available. Up to now, the system has been build for Czech, English, Russian and German.[5] In this paper, we focus only on its implementation for Czech and English.

3.1 Preprocessing to a Tectogrammatical Representation

Before coreference resolution is carried out, the input text must undergo a thorough analysis producing a tectogrammatical representation of its sentences. Treex CR cannot process a text that has not been analyzed this way. Input data must comply with at least basics of this annotation style. The text should be tokenized and labeled with part-of-speech tags in order for the resolver to focus on nouns and pronouns as mention candidates. However, the real power of the system lies in exploiting rich linguistic annotation that can be represented by tectogramatics.

Czech and English analysis. We make use of rich pipelines for Czech and English available in the Treex framework, previously applied for building the Czech-English parallel treebank CzEng 1.6 [4].

Sentences are first split into tokens, which is ensured by rule-based modules. Subsequently, the tokens are enriched with morphological information including part-of-speech tag, morphological features as well as lemmas. Whereas in English, the Morče tool [33] is used to collect part-of-speech tags, followed by a rule-based lemmatizer, the Czech pipeline utilizes the MorphoDiTa tool [34] to obtain all.

A dependency tree is build on top of this annotation, using MST parser [19] and its adapted version [28] for English and Czech, respectively. Named entity recognition is carried out by the NameTag [35] tool in both languages.

The NADA tool [3] is applied to help distinguish referential and non-referential occurrences of the English pronoun "*it*". Every occurrence is assigned a probability estimate based on n-gram features.

Transition from a surface dependency tree to the tectogrammatical one consists of the following steps. As tectogrammatical nodes correspond to content words only, function words such as prepositions, auxiliary verbs, particles, punctuation must be hidden. Morpho-syntactic information is transferred to tectogrammatical layer by two channels: (i) morpho-syntactic tags called *formemes* [13] and (ii) features of deep grammar called *grammatemes*. All nodes are then subject to semantic role labeling assigning them roles such as Actor and Patient, and linking of verbs to items in Czech valency dictionary [12].

[5] Russian and German version has been trained on automatic English coreference labeling projected to these languages through a parallel corpus. See [27] for more details.

Reconstructing zeros. To mimic the style of tectogrammatical annotation in automatic analysis, some nodes that are not present on the surface must be reconstructed. We focus on cases that directly relate to coreference. Such nodes are added by heuristics based on syntactic structures.

Subject zeros are the most prominent anaphoric zeros in Czech. A subject is generated as a child of a finite verb if it has no children in subject position or in nominative case. Grammatical person, number and gender are inferred from a form of the verb.

Perhaps surprisingly, English uses zeros as well. The coreferential ones can be found in *relative clauses* (see the example in Section 1) and *non-finite verbal constructions*, e.g. in participles and infinitives. We seek for such constructions and add a zero child with a semantic role corresponding to the type of the construction. This work extends the original Treex module for English zeros' generation, which addressed only infinitives.

3.2 Model design

Treex CR models coreference in a way to be easily optimized by supervised learning. Particularly, we use logistic regression with stochastic gradient descend optimization implemented in the Vowpal Wabbit toolkit.[6] Design of the model employs multiple concepts that have proved to be useful and simple at the same time.

Mention-ranking model. Given an anaphor and a set of antecedent candidates, *mention-ranking* models [10] are trained to score all the candidates at once. Competition between the candidates is captured in the model. Every antecedent candidate describes solely the actual mention. It does not represent a possible cluster of coreferential mentions built up to the moment.

Antecedent candidates for an anaphor are selected from the context window of a predefined size. This is done only for the nodes satisfying simple morphological criteria (e.g. nouns and pronouns). Both the window size and the filtering criteria can be altered as hyperparameters.

Joint anaphoricity detection and antecedent selection. What we denote as an anaphor in the model is, in fact, an anaphor candidate. There is no preprocessing that would filter out non-referential anaphor candidates. Instead, both decisions, i.e. (i) to determine if the anaphor candidate is referential and (ii) to find the antecedent of the anaphor, are performed in a single step. This is ensured by adding a fake "antecedent" candidate representing solely the anaphor candidate itself. By selecting this candidate, the model labels the anaphor candidate as non-referential.

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

A cascade of specialized models. Properties of coreferential relations are so diverse that it is worth modeling individual anaphor types rather separately than jointly as shown in [11]. For instance, while personal pronouns may refer to one of the previous sentences, the antecedent of relative and reflexive pronouns always lies in the same sentence. By representing coreference of these expressions separately in multiple specialized models, the abovementioned hyperparameters can be adjusted to suit the particular anaphor type. Processing of these anaphor types may be sorted in a cascade so that the output of one model might be taken into account in the following models. Currently, we do not take advantage of this feature, though. Models are thus independent on each other and can be run in any ordering.

3.3 Feature extraction

The preprocessing stage (see Section 3.1) enriches a raw text with substantial amount of linguistic material. Feature extraction stage then uses this material to yield *features* consumable by the learning method. In addition, Vowpal Wabbit, the learning tool we use, supports grouping features into namespaces. The tool may introduce combinations of features as a Cartesian product of selected namespaces and thus massively extend the space of features. This can be controlled by hyperparameters to Vowpal Wabbit.

Features used in Treex CR can be categorized by their form. The categories differ in the number of input arguments they require. *Unary features* describe only a single node, either anaphor or antecedent candidate. Such features start with prefixes `anaph` and `cand`, respectively. *Binary features* require both the anaphor and the antecedent candidate for their construction. Specifically, they can be formed by agreement or concatenation of respective unary features, but they can generally describe any relation between the two arguments. Finally, *ranking features* need all the antecedent candidates along with the anaphor candidate to be yielded. Their purpose is to rank antecedent candidates with respect to a particular relation to an anaphor candidate.

Our features also differ by their content. They can be divided into three categories: (i) location and distance features, (ii) (deep) morpho-syntactic features, and (iii) lexical features. The core of the feature set was formed by adapting features introduced in [25].

Location and distance features Positions of anaphor and an antecedent in a sentence were inspired by [6]. Position of the antecedent is measured backward from the anaphor if they lie in the same sentence, otherwise it is measured forward from the start of the sentence. As for distance features, we use various granularity to measure distance between an anaphor and an antecedent candidate: number of sentences, clauses and words. In addition, an ordinal number of the current candidate antecedent among the others is

included. All location and distance features are bucketed into predefined bins.

(Deep) morpho-syntactic features. They utilize the annotation provided by part-of-speech taggers, parsers and tectogrammatical annotation. Their unary variants capture the mention head's part-of-speech tag and morphological features, e.g. gender, number, person, case. As gender and number are considered important for resolution of pronouns, we do not rely on their disambiguation and work with all possible hypotheses. We do the same for some Czech words that are in nominative case but disambiguation labeled them with the accusative case. Such case is a typical source of errors in generating a subject zero as it fills a missing nominative slot in the governing verb's valency frame. To discover potentially spurious subject zeros, we also inspect if the verb has multiple arguments in accusative and if the argument in nominative is refused by the valency, as it is in the phrase "*Zdá se mi, že...*" ("*It seems to me that...*"). Furthermore, the unary features contain (deep) syntax features including its dependency relation, semantic role, and formeme. We exploit the structure of the syntactic tree as well, extracting some features from the mention head's parent.

Many of these features are combined to binary variants by agreement and concatenation. Heuristics used in original Treex modules for some anaphor types gave birth to another pack of binary features. For instance, the feature indicating if a candidate is the subject of the anaphor's clause should target coreference of reflexive pronouns. Similarly, signaling whether a candidate governs the anaphor's clause should help with resolution of relative pronouns.

Lexical features Lemmas of the mentions' heads and their parents are directly used as features. Such features may have an effect only if built from frequent words, though. By using them with an external lexical resource, this data sparsity problem can be reduced.

Firstly, we used a long list of noun-verb collocations collected by [25] on Czech National Corpus [9]. Having this statistics, we can estimate how probable is that the anaphor's governing verb collocates with an antecedent candidate.

Another approach to fight data sparsity is to employ an ontology. Apart from an actual word, we can include all its hypernymous concepts from the hierarchy as features. We exploit WordNet [14] and EuroWordNet [38] for English and Czech, respectively.

To target proper nouns, we also extract features from tags assigned by named entity recognizers run during the preprocessing stage.

4 Datasets

We exploited two treebanks for training and testing purposes: Prague Dependency Treebank 3.0 [2, PDT] and Prague Czech-English Dependency Treebank 2.0 Coref [22, PCEDT] for Czech and English, respectively. Although PCEDT is a Czech-English parallel treebank, we used only its English side. Both treebanks are collections of newspaper and journal articles. In addition, they both follow the annotation principles of the theory of Prague tectogrammatics [32]. They also comprise a full-fledged manual annotation of coreferential relations.[7]

	Czech		English		
	Train	Eval test	Train	Eval test	CoNLL 2012
sents	38k	5k	39k	5k	9.5k
words	652k	92k	912k	130k	170k
t-nodes	528k	75k	652k	91k	116k
anaph	92k	14k	103k	15k	15k
Relative	7.2k	1k	6.4k	0.8k	–
Reflexive	3.4k	0.6k	0.4k	0.05k	0.1k
PP3	–	–	19k	2.4k	4.5k
SzPP3	12k	2k	–	–	–
Zero	–	–	23k	3.2k	–
Other	70k	10k	54k	8.0k	10.4k

Table 1: Basic statistics of used datasets. The class *SzPP3* stands for 3rd person subject zeros, personal and possessive pronouns, while the class *PP3* excludes subject zeros.

Training and *evaluation test* dataset for Czech are formed by PDT sections `train-*` and `etest`, respectively. As for English, these two datasets are collected from PCEDT sections `00-18` and `22-24`, respectively.[8]

In addition, we used the official testset for CoNLL 2012 Shared Task to evaluate English systems [31]. This dataset has been sampled from the OntoNotes 5.0 [30] corpus. OntoNotes, and thus CoNLL 2012 testset as well, differ from the two treebanks in the following main aspects: (i) coreference is annotated on the surface, where mentions of the same entity are co-indexed spans of consecutive words, (ii) it contains no zeros and relative pronouns are not annotated for coreference.[9] These differences must be reflected when evaluating on this dataset (see Section 5).

A basic statistics collected on these datasets is shown in Table 1. The anaphor types treated by Treex CR cover around 50% and 25-30% of all anaphors in English and Czech tectogrammatical treebanks, respectively. The main reason of the disproportion is that we did not include Czech non-subject zeros to the collection (class *Zero*). Czech subject zeros are merged to a common class with personal and possessive pronouns in 3rd person (class *SzPP3*), as they are trained in a joint model (see Section 5). Due the same reason, English personal and possessive pronouns in 3rd person form a common class *PP3*. As the CoNLL 2012 testset has no annotation of relative pronouns and zeros, Treex CR covers 30% of all the anaphors.

[7] See [21] for more information on coreference annotation.

[8] During development of our system, we employed the rest of the treebanks' data as *development test* dataset for intermediate testing.

[9] Reasons for ignoring relative pronouns in OntoNotes are unclear. They might be seen as so tied up with rules of grammar and syntax that annotation of such cases is too unattractive to deal with.

5 Experiments and Evaluation

Our system uses two specialized models for relative and reflexive pronouns in both languages. The Czech system in addition contains a joint model for subject zeros, personal and possessive pronouns in 3rd person (denoted as *SzPP3*). The English system contains two more models: one for personal and possessive pronouns in 3rd person (denoted as *PP3*) and another one for zeros.

Systems to compare. To show performance of Treex CR in a context, we evaluated multiple other systems on the same data. Since currently there is no other publicly available system for Czech to our knowledge, we compare it with the original Treex set of modules for coreference. The set consists of rule-based modules for relative and reflexive pronouns, and a supervised model for SzPP3 mentions. It has been previously used for building a Czech-English parallel treebank CzEng 1.0 [5].

We also report performance of the English predecessor of Treex CR used to build CzEng 1.0. It comprises a rule-based module for relative pronouns and zeros, and a joint supervised model for reflexives and PP3 mentions. In addition, we include the Stanford Core NLP toolkit to the evaluation. It contains three approaches to full-fledged CR that all claimed to improve over the state of the art at the time of their release: deterministic [18], statistical [7], and neural [8]. In fact, the neural system has not been outperformed, yet.

Stanford Core NLP predicts surface mentions, which is not compatible with the evaluation schema designed for tectogrammatical trees. The surface mentions thus must be transformed to the tectogrammatical style of coreference annotation, i.e. the mention heads must be connected with links. We may use the information on mention heads provided by the Stanford system itself. However, by using this approach results we observed completely contradictory results on different datasets. Manual investigation on a sample of the data revealed that often the Stanford system in fact identified a correct antecedent mention, but selected a head different to the one in the data. Most of these cases, e.g. company names like "McDonald's Corp." or "Walt Disney Co.", have no clear head, though. Therefore, we decided to use the gold tectogrammatical tree to identify the head of the mention labeled by the Stanford system. Even though employing gold information for system's decision is a bad practice, here it should not affect the result so much and we use it only for the third-party systems, not for our Treex CR.

Evaluation measure. Standard evaluation metrics (e.g. MUC [37], B^3 [1]) are not suitable for our purposes as they do not allow for scoring only a subset of mentions. Instead, we use a measure similar to scores proposed by [36]. For an anaphor candidate a_i, we increment the three following counts:

- $true(a_i)$ if a_i is anaphoric in the gold annotation,
- $pred(a_i)$ if the CR system claims a_i is anaphoric,
- $both(a_i)$ if both the system and gold annotation claim a_i is anaphoric and the antecedent found by the system belongs to the transitive closure of all mentions coreferential with a_i in the gold annotation.

After aggregating these counts over all anaphor candidates, we compute the final Precision, Recall and F-score as follows:

$$P = \sum_{a_i} \frac{both(a_i)}{pred(a_i)} \quad R = \sum_{a_i} \frac{both(a_i)}{true(a_i)} \quad F = \frac{2PR}{P+R}$$

To evaluate only a particular anaphor type, the aggregation runs over all anaphor candidates of the given type.

The presented evaluation schema, however, needs to be adjusted for the CoNLL 2012 dataset. As mentioned in Section 4, in this dataset relative pronouns are not considered coreferential and zeros are missing at all. As a result, a system that marks such expressions as antecedents would be penalized. We thus apply the following patch specifically for the CoNLL 2012 dataset to rectify this issue. If the predicted antecedent is a zero or a relative pronoun, instead of using it directly we follow the predicted coreferential chain until the expression outside of these two categories is met. The found expression is then used to calculate the counts, as described above. If no such expression is found, the direct antecedent is used, even if it is a zero or a relative pronoun.

All the scores presented in the rest of the paper are F-scores.

	Relative	Reflexive	SzPP3	All
Count	1,075	579	1,950	3,604
Treex				
CzEng 1.0	57.14	67.57	50.52	55.20
Treex CR	78.40	76.19	61.31	**68.46**

Table 2: F-scores of Czech coreference resolvers measured on all anaphor types both separately and altogether. The type *SzPP3* denotes 3rd person subject zeros, personal and possessive pronouns.

Results and their analysis. Table 2 shows results of evaluation on the Czech data. The Czech version of Treex CR succeeded in its ambition to replace the modules used in Treex until now. It significantly[10] outperformed the baseline for each of the anaphor type, with the overall score by 13 percentage points higher. The jump for relative pronouns was particularly high.

The analysis of improved examples for this category shows that apart from the syntactic principles used in the rule-based module, it also exploits other symptoms of

[10] Significance has been calculated by bootstrap resampling with a confidence level 95%.

	PCEDT Eval					CoNLL 2012 test set		
	Relative	Reflexive	PP3	Zeros	All	Reflexive	PP3	All
Count	842	49	2,494	3,260	6,645	111	4,583	4,710
Stanford								
deterministic	1.16	55.67	63.65	0.00	34.96	71.11	60.55	60.79
statistical	0.00	63.74	72.71	0.00	39.09	80.56	*71.07*	**71.29**
neural	0.00	70.97	*76.36*	0.00	41.56	*80.73*	70.45	70.70
Treex								
CzEng 1.0	70.64	65.93	73.52	28.48	55.34	76.02	67.93	68.13
Treex CR	*75.99*	*81.63*	74.11	*45.37*	**60.87**	79.65	66.64	66.96

Table 3: F-scores of English coreference resolvers measured on all anaphor types both separately and altogether. The type *PP3* denotes personal and possessive pronouns in 3rd person.

coreference. The most prominent are agreement of the anaphor and the antecedent in gender and number as well as the distance between the two. It also succeeds in identifying non-anaphoric examples, for instance interrogative pronouns, which use the same forms.

Results of evaluation on the English data are highlighted in Table 3. Similarly to the Czech system, the English version of Treex CR outperforms its predecessor in Treex by a large margin of 15 percentage points on the PCEDT Eval testset. Most of it stems from a large improvement on the biggest class of anaphors, zeros. Unlike for Czech relative pronouns, the supervised CR is not the only reason for this leap. It largely results from the extension that we made to the method for adding zero arguments of non-finite clauses (see Section 3.1). Consequently, the coverage of these nodes compared to their gold annotation rose from 34% to 80%. Comparing these two versions of the Treex system on the CoNLL 2012 testset, we see a different picture. The systems' performances are more similar, the baseline system for PP3 even slightly outperforms the new Treex CR.

As for the comparison with the Stanford systems, we should not look at the scores aggregated over all the anaphor types under scrutiny, because Stanford systems apparently do not address zeros and relative pronouns.[11] In fact, the Stanford systems try to reconstruct coreference as it is annotated in OntoNotes 5.0.

The classes of reflexive and PP3 pronouns are the only ones within the scope of all the resolvers. The Stanford deterministic system seems to be consistently outperformed by all the other approaches. Performance rankings on reflexive pronouns differ for the two datasets, which is probably the consequence of low frequency of reflexives in the datasets. Regarding the PP3 pronouns, Treex CR does not achieve the performance of the state-of-the-art Stanford neural system. On the CoNLL 2012 testset it is outperformed even by the Stanford statistical system. Nevertheless, in all the cases the performance gaps are not so big and thus it is reasonable using Treex CR for further experiments in the future.

To best of our knowledge, no analysis of how Stanford systems perform for individual anaphor types has been published, yet. Interestingly, our result show that even though the overall performance of the neural system on the CoNLL 2012 testset is reported to be higher [8], for personal and possessive pronouns in third person it is slightly outperformed by the statistical system. However, as the evaluation on the PCEDT Eval testset shows completely the opposite, we cannot arrive at any conclusion on their mutual performance comparison on this anaphor type.

6 Conclusion

We described Treex CR, a coreference resolver not only for Czech. The main feature of the system is that it operates on the tectogrammatical layer, which allows it to address also coreference of zeros. The system uses a supervised model, supported by a very rich set of linguistic features. We presented modules for processing Czech and English and evaluated them on several datasets. For comparison, we conducted the evaluation with the predecessors of Treex CR and three versions of the Stanford system, one of which was a state-of-the-art neural resolver for English. Our system seems to have outperformed the baseline system on Czech. On English, although it could not outperform the best approaches in the Stanford system, its performance is high enough to be used in future experiments. Furthermore, it may be used for resolution of anaphor types that are ignored by most of the coreference resolvers for English, i.e. relative pronouns and zeros.

In the future work, we would like to use Treex CR in cross-lingual coreference resolution, where the system is applied on parallel corpus and thus it may take advantage of both languages.

[11] On the other hand, they address coreference of nominal groups and pronouns in first and second person. Treex CR does not provide Czech or English models for these classes, so far. Nevertheless, experimental projection-based models already exist for German and Russian [27].

Acknowledgments

This project has been funded by the GAUK grant 338915 and the Czech Science Foundation grant GA-16-05394S. This work has been also supported and has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project No. LM2015071 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Amit Bagga and Breck Baldwin. Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566, 1998.

[2] Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013.

[3] Shane Bergsma and David Yarowsky. NADA: A Robust System for Non-referential Pronoun Detection. In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, pages 12–23, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Artificial Intelligence, pages 231–238, Heidelberg, Germany, 2016. Springer International Publishing.

[5] Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3921–3928, Istanbul, Turkey, 2012. European Language Resources Association.

[6] Eugene Charniak and Micha Elsner. EM Works for Pronoun Anaphora Resolution. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 148–156, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[7] Kevin Clark and Christopher D. Manning. Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, 2015. Association for Computational Linguistics.

[8] Kevin Clark and Christopher D. Manning. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, 2016. Association for Computational Linguistics.

[9] CNC. Czech national corpus – SYN2005, 2005.

[10] Pascal Denis and Jason Baldridge. A Ranking Approach to Pronoun Resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1588–1593, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[11] Pascal Denis and Jason Baldridge. Specialized Models and Ranking for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[12] Ondřej Dušek, Jan Hajič, and Zdeňka Urešová. Verbal Valency Frame Detection and Selection in Czech and English. In *The 2nd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[13] Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. Formemes in English-Czech Deep Syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 267–274, Montréal, Canada, 2012. Association for Computational Linguistics.

[14] Christiane Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[15] Aria Haghighi and Dan Klein. Coreference Resolution in a Modular, Entity-centered Model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[16] Jan Hajič et al. Prague Dependency Treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

[17] Lucie Kučová and Zdeněk Žabokrtský. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora. In *Lecture Notes in Artificial Intelligence, Proceedings of the 8th International Conference, TSD 2005*, volume 3658 of *Lecture Notes in Computer Science*, pages 93–98, Berlin / Heidelberg, 2005. Springer.

[18] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[19] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective Dependency Parsing Using Spanning Tree Algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[20] Ruslan Mitkov. *Anaphora Resolution*. Longman, London, 2002.

[21] Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 169–176, Paris, France, 2016. European Language Resources Association.

[22] Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. Prague czech-english dependency treebank 2.0 coref, 2016.

[23] Vincent Ng. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[24] Giang Linh Nguy. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, MFF UK, Prague, Czech Republic, 2006. In Czech.

[25] Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, pages 276–285, London, UK, 2009. The Association for Computational Linguistics.

[26] Michal Novák and Anna Nedoluzhko. Correspondences between Czech and English Coreferential Expressions. *Discours: Revue de linguistique, psycholinguistique et informatique.*, 16:1–41, 2015.

[27] Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. Projection-based Coreference Resolution Using Deep Syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Valencia, Spain, 2017. Association for Computational Linguistics.

[28] Václav Novák and Zdeněk Žabokrtský. Feature engineering in maximum spanning tree dependency parser. volume 4629, pages 92–98, Berlin / Heidelberg, 2007. Springer.

[29] Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP Framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

[30] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards Robust Linguistic Analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[31] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012*, pages 1–40, Jeju, Korea, 2012. Association for Computational Linguistics.

[32] Petr Sgall, Eva Hajičová, Jarmila Panevová, and Jacob Mey. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer, 1986.

[33] Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The Best of Two Worlds: Cooperation of Statistical and Rule-based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, pages 67–74, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[34] Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[35] Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, 2014. Association for Computational Linguistics.

[36] Don Tuggener. Coreference Resolution Evaluation for Higher Level Applications. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 231–235. The Association for Computer Linguistics, 2014.

[37] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A Model-theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[38] Piek Vossen. Introduction to EuroWordNet. *Computers and the Humanities, Special Issue on EuroWordNet*, 32(2–3), 1998.