Pronoun Prediction with Linguistic Features and Example Weighing

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

Abstract

We present a system submitted to the WMT16 shared task in cross-lingual pronoun prediction, in particular, to the English-to-German and German-to-English sub-tasks. The system is based on a linear classifier making use of features both from the target language model and from linguistically analyzed source and target texts. Furthermore, we apply example weighing in classifier learning, which proved to be beneficial for recall in less frequent pronoun classes. Compared to other shared task participants, our best English-to-German system is able to rank just below the top performing submissions.

1 Introduction

Previous works concerning translation of pronouns[1] have shown that unlike other words, pronouns require a special treatment. Context and target language grammar influence pronoun translation much more profoundly than the translation of parts-of-speech carrying lexical information.

This paper presents a system for the WMT16 shared task of cross-lingual pronoun prediction (Guillou et al., 2016),[2] the task that looks at the problem of pronoun translation in a more simplified way. Here, the objective is to predict a target language pronoun from a set of possible candidates, given source text, lemmatized and part-of-speech-tagged target text, and automatic word alignment. We address specifically the sub-tasks of English-to-German and German-to-English pronoun prediction.

We take a machine learning approach to the problem and apply a linear classifier. Our approach combines features coming from the target language model with features extracted from the linguistically analyzed source and target texts. We also introduce training example weighing, which aims at improving the prediction accuracy of less populated target pronouns. All the source codes used to build the system are publicly available.[3]

According to the WMT16 pronoun translation shared task results (Guillou et al., 2016), our best German-to-English system ranks in the middle of the pack while our English-to-German systems seem to be the poorest. However, after the shared task submission deadline, we discovered an error in post-processing of the classifier predictions on the evaluation set for the English-to-German direction. After correcting this error, our system reaches the 2nd best result for this language direction.

The paper is structured as follows. After introducing the related work in Section 7, we describe three preprocessing components of our system that enrich the input data with additional information in Section 2. Section 3 then presents features extracted from the data whereas Section 4 gives more details about the method used to train the model. In Section 5, all our system configurations submitted to the shared tasks are evaluated. Finally, we examine the effect of individual features and example weighing in Section 6 before we conclude in Section 8.

2 Preprocessing components

The preprocessing stage combines three components, each of them enriching the input data with additional information: a target language model, an automatic linguistic analysis of the source sen-

[1] Summarized by Hardmeier (2014).

[2] http://www.statmt.org/wmt16/pronoun-task.html

[3] https://github.com/ufal/wmt16-pronouns

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 602–608,
Berlin, Germany, August 11-12, 2016.

tences, and a basic automatic analysis of the target sentences.

2.1 Target language model

For language modeling, we employed the KenLM Language Model Toolkit (Heafield et al., 2013), an efficient implementation of large language models with modified Kneser-Ney smoothing (Kneser and Ney, 1995).

Lemmatized 5-gram models for English and German have been supplied as a baseline system by the organizers of the shared task. An integral part of the baseline system is a wrapper script[4] performing necessary preprocessing before the actual probability estimation. For instance, it selects words which may possibly belong to the `OTHER` class[5] and it enables setting a penalty for preferring an empty word.[6] We only adjusted the wrapper script so that it fits into our processing pipeline, making no modifications to the estimation machinery.

2.2 Source language analysis

In the input data supplied by the task organizers, source text is represented as plain tokenized sentences. We have processed the source texts with tools obtaining additional linguistic analysis. However, due to different availability of these tools for English and German, the depth of the analysis differs. We describe both analysis pipelines separately in the following:

English. English source texts have been analyzed up to the level of deep syntax using the Treex framework (Popel and Žabokrtský, 2010) incorporating several external tools. The processing pipeline consists of part-of-speech tagging with the Morče tool (Spoustová et al., 2007) dependency parsing conducted by the MST parser (McDonald et al., 2005), semantic role labeling (Bojar et al., 2016), and coreference resolution obtained as a combination of Treex coreference modules and the Bart 2 toolkit (Versley et al., 2008; Uryupina et al., 2012). Prior to the last step, all instances of the pronoun *it* are assigned a probability of being anaphoric by the NADA tool (Bergsma and Yarowsky, 2011).

German. We utilized the MATE tools[7] (Björkelund et al., 2010) to perform part-of-speech tagging, morphological analysis (necessary to obtain grammatical categories such as gender or number), and transition-based dependency parsing (Bohnet and Nivre, 2012; Seeker and Kuhn, 2012).

2.3 Target language analysis

In the data supplied by the task organizers, the format of the target language sentences differs from the source language format. Not only are the target words to be predicted replaced by a placeholder, but all other tokens are also substituted with corresponding lemmas and coarse-grained part-of-speech tags.

For this reason, we needed to simplify the analysis of target texts. The parsers used for source texts do not accept the tagset used by the organizers. There are two possible solutions to fix this disagreement: either running a part-of-speech tagger producing tags that agree with the tagset required by the parser, or obtaining suitable part-of-speech tags by a transformation of the original tagset. However, both options are prone to errors. In the former option, the tags produced in this way would definitely be of low quality as only a lemmatized text is available. This would cause problems especially for German. The latter option brings another problem. The original tagsets (12 tags in both English and German) are more coarse-grained than the tagsets required by the parsers (44 and 53 tags in English and German, respectively), which makes the transformation in this direction difficult.

Due to these obstacles, we decided to abandon any additional linguistic processing except for the identification of noun genders. We consider gender and number information one of the most valuable inputs for correct pronoun translation. While the number information is hard to reconstruct from a lemmatized text with part-of-speech tags having no indication of grammatical number, gender can be reconstructed from a noun lemma itself quite satisfactorily. In each of the languages, we approached the task of obtaining gender for a given noun in a different way.

[4] `https://bitbucket.org/yannick/discomt_baseline/src`

[5] The `OTHER` class comprise words, not necessarily pronouns, that appear often enough in the context typical for pronouns to be resolved but not enough to form their own class. Furthermore, it can be an empty word if the source pronoun has no target language counterpart.

[6] In all experiments, we used zero penalty.

[7] `https://code.google.com/archive/p/mate-tools/`

English. The gender information was obtained using the data collected by Bergsma and Lin (2006).[8] They used paths in dependency trees to learn the likelihood of coreference between a pronoun and a noun candidate and then applied them in a bootstrapping fashion on larger data to obtain a noun gender and number distribution in different contexts.

For the sake of simplicity, we filtered their list only to single-word items. If we encounter a token with a noun tag assigned in the target sentence, its lemma is looked up in the list and assigned the most probable gender, if any is found. Otherwise, the neuter gender is assumed.

German. We run the MATE morphological analysis separately for every lemma labeled as a noun. If no gender information is obtained, the noun is assigned the neuter gender.

3 Feature extraction

Having both the source and the target texts enriched with additional linguistic information, we extract a set of instances that are later fed into into our classifier. An instance is extracted for every target-language pronoun (placeholder) to be classified represented by features that can be divided into several categories:

Target language model features. Using the KenLM with the wrapper supplied by the organizers, we obtain an estimated probability value for every candidate pronoun. From this, we produce features describing the actual probability values for each candidate word, quantized into 9 bins. Furthermore, features ranking the candidate words by their probabilities, quantized in three different ways, are extracted.

Source language features. The data supplied by the organizers also contain automatic word alignment between the source and the target sentences. Therefore, when extracting features for a given placeholder in the target language, we are able to do the same for its counterparts in the source language. Deeper linguistic analysis performed for the source language (see Section 2.2) allows us to extract richer features than for the target language.

For every source counterpart of a target pronoun placeholder, we extract its lemma, syntactic dependency function, the lemma of its parent in the dependency tree, and combinations of the previous features. As the analysis of English goes deeper than the surface syntax, we include the semantic function of the source counterpart. If the counterpart is an instance of the pronoun *it*, we add the anaphoricity probability estimated by the NADA detector, quantized in the same way as the probabilities coming from the KenLM model.

Target language features. The lemma of a parent verb of the target pronoun placeholder might also be a valuable feature. Even though we have not performed a syntactic analysis on the target text (see Section 2.3), we are still able to approximate it in several ways. The easiest option is to list all verb lemmas that appear in a relatively small context surrounding the placeholder (1, 3, or 5 words). Another approach is to project the parent dependency relation from the source sentence via word alignment. We also extract the part-of-speech tags of the parents collected in this way, since they might not be verbs due to possible errors.

Antecedent features. The gender of an anaphoric pronoun is often determined by the gender of its antecedent. Same as with syntactic trees, we have no information on coreference in the target text. Again, we approximate it in two ways. We project the coreference link via word alignment and use the gender of the projected antecedent. Note that this approach can be used only in the English-to-German direction due to missing coreference resolution for German. To extract similar information also for the opposite direction, we take advantage of the fact that the task is defined for subject pronouns only. A tendency of consecutive subjects to refer to the same entity inspired us to include the gender of the previous target language subject as a feature. The indicator whether a word is a subject is again projected via alignment from the source text.

4 Model

Pronoun prediction as specified by the organizers is a classification task. We address it by machine learning, building a linear model using the multiclass variant of the logistic loss and stochastic gradient descent optimization as implemented in the Vowpal Wabbit toolkit.[9] To train the model, we

[8] `http://www.clsp.jhu.edu/~sbergsma/Gender/Data/`

[9] Available at `https://github.com/JohnLangford/vowpal_wabbit/wiki`. Vowpal Wabbit has been chosen due

	Name	Setting	Dev MACRO-R	Dev ACC	Eval MACRO-R	Eval ACC
EN-to-DE	baseline	—	34.35	42.81	38.53	50.13
	CUNI-primary	weighted, passes: 5, L1: 3×10^{-7}	**45.63**	57.72	*__54.37__	*64.23
	CUNI-contrastive	unweighted, passes: 1, L1: 5×10^{-6}	42.54	63.51	*51.74	*71.80
DE-to-EN	baseline	—	36.08	50.47	42.15	53.42
	CUNI-primary	weighted, passes: 1, L1: 0	**56.47**	68.35	**60.42**	64.18
	CUNI-contrastive	unweighted, passes: 5, L1: 0	51.62	70.59	56.83	65.22

Table 1: Our Systems submitted to the shared task and their performance compared to the baseline system. The official measure of performance is macro-averaged recall (MACRO-R), while accuracy (ACC) serves as a contrastive measure. Scores labeled by the * symbol differ from the official results of the shared task (Guillou et al., 2016) as an error has been discovered after the task submission deadline.

run the learner over the training data with features described in Section 3, possibly in multiple passes and with various rates of L1 or L2 regularization.

Optimization with respect to the logistic loss function is a widely used approximation of the the accuracy measure. However, the official scoring metric set by the task organizers is the macro-averaged recall. Macro-averaging causes that improvements in recall for less frequent target pronouns have a stronger effect than improvements for more frequent pronouns. We address this issue by weighing the training data instances based on the target class. We weigh the classes in an inverse proportion to how frequently they appear in the training data. The less frequent a pronoun is, the heavier penalty is incurred if it is misclassified.

5 Submitted systems

We submitted four systems to the shared task – two systems to each of the two sub-tasks: English-to-German and German-to-English prediction. The systems trained on the weighted examples are considered as *primary* while the unweighted systems were submitted as *contrastive*.

Training examples have been extracted from all the data supplied for training by the organizers.[10] The same holds for the data designated for development and evaluation testing.

The best combination of learning parameters has been selected by a grid search with various parameter settings on the development data. Table 1 specifies the learning parameters used for all systems submitted. It also shows macro-averaged recall and accuracy measured on both the development and the evaluation set Moreover, it and compares the performance with the baseline system based on the KenLM target language model as supplied by the organizers (see Section 2.1).

Note that the scores of our English-to-German systems achieved on the evaluation set are much better than the scores presented in the official results of the shared task Guillou et al. (2016). An error that concerned merging of the classifier output into the test data file for submission, which was, however, discovered after the deadline for task submissions. According to the official results, our German-to-English primary system is ranked fourth among six participating primary systems. Our English-to-German primary system, ranked last among nine systems in the official results, would place as second if we took the correct scores.

6 Feature ablation and weighing analysis

In order to assess the effect of individual feature types, we carried out an additional experiment. For both translation directions we trained models on various subsets of the complete feature set. All the models have been trained in both weighted and unweighted scenarios.

The experiments were conducted with the following feature sets:

- `all`: the complete feature set as described in Section 3
- `-src`: the complete feature set, excluding source language features

to its fastest throughput among all machine learning tools known to us as well as due to the remarkable variety of options for learning, e.g. example weighing used in our experiments. However, there are still options that are worth to be examined in future experiments, for instance using other loss functions, e.g. a hinge loss which is equivalent to the SVM algorithm.

[10] `http://data.statmt.org/wmt16/pronoun-task/`

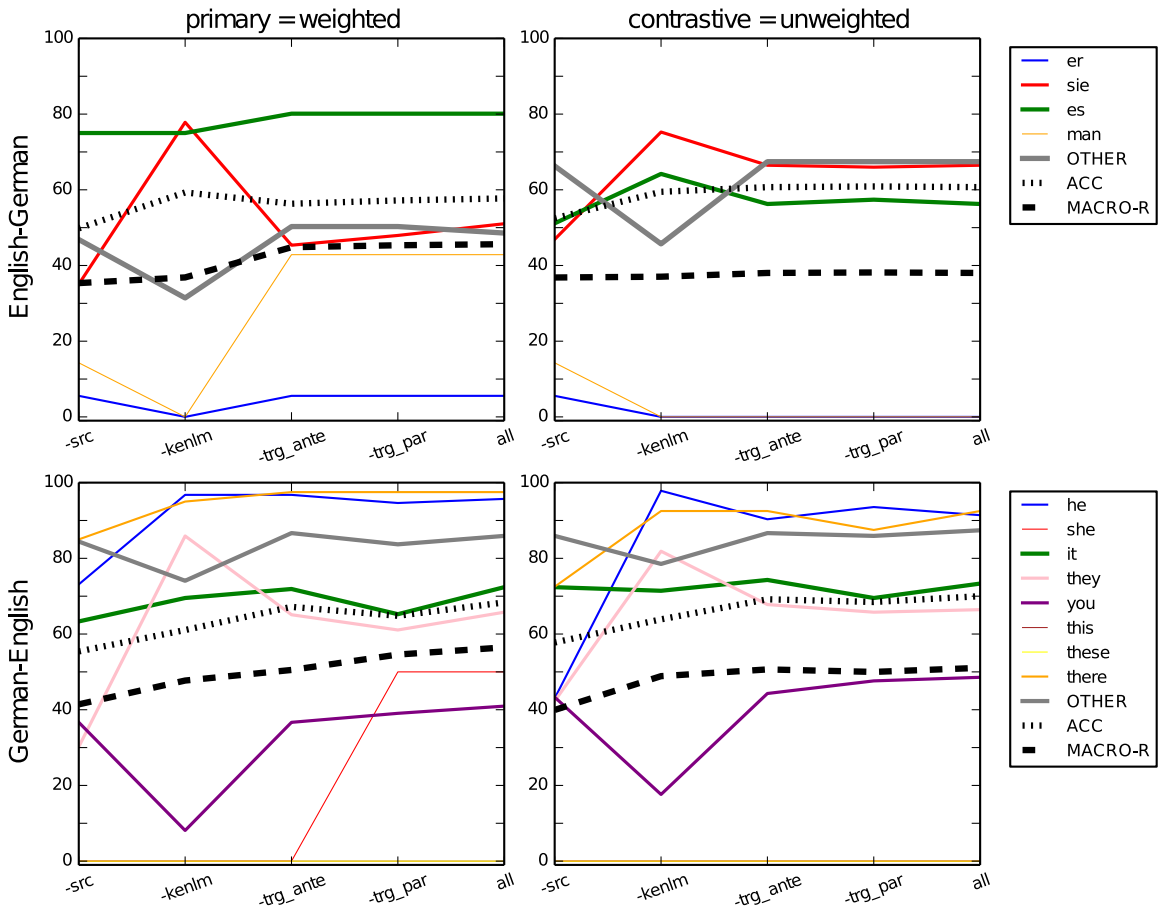

Figure 1: The impact of feature ablation on per-class recall (see Section 6 for details), macro-averaged recall (MACRO-R), and accuracy (ACC) in the four systems submitted to the shared task.

- `-kenlm`: the complete feature set, excluding KenLM features
- `-trg_ante`: the complete feature set, excluding features approximating the gender of the antecedent of the target pronoun
- `-trg_par`: the complete feature set, excluding features approximating the parent of the target pronoun

Figure 1 shows the performance of weighted and unweighted models for both translation directions if trained in all of the feature settings listed above. The performance is measured by recall on each of the target classes (solid color lines, whose widths illustrate the frequency of the class in the training data), as well as by micro-averaged recall, which equals to overall accuracy for this task (dotted line), and macro-averaged recall, which is the official measure in the shared task (dashed line).

The graphs show that the impact of individual feature categories on the macro-averaged recall is generally higher in the German-English direction and for weighted models. For instance, leaving out the most valuable category of source language features decreases the performance level by just 1 percentage point for the English-German unweighted model while degrading the performance of the German-English model by 15 percentage points. The graphs also show that the KenLM features have the strongest effect on the final recall values for the individual pronoun classes. A positive effect of English coreference resolution to determining the correct gender of a German pronoun can be also observed. Adding antecedent features to English-to-German weighted system causes a small recall increase of the pronoun *sie* with almost no degradation to other classes.

The impact of instance weighing turns out to be more interesting. Focusing on scores for individual classes, one can observe that the pronouns that benefit from weighing the most are the less frequent ones, i.e., *man* and *er* in German, *there*, *he*, and *she* in English. On the other hand, the effect of weighing reduces performance in frequent

classes, such as *OTHER*, German *sie*, and English *you*. The only exception is the German pronoun *es*, whose recall rises for weighted models even though it is one of the most frequent pronoun classes. Overall, instance weighting fulfills our expectations: although it causes a decrease in recall for frequent pronoun classes, it improves the official macro-averaged recall score.

7 Related work

A similar problem was addressed in the DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) as a cross-lingual pronoun prediction subtask. It differed from the current task in one main aspect: the manually translated target text was available in its surface form as an input, i.e., it was neither machine-translated nor lemmatized and part-of-speech-tagged at least as it is in the WMT16 shared task. This aspect, far from a real-world machine translation scenario, probably caused that none of the participants was able to beat the baseline, the target language model.

Out of the DiscoMT 2015 shared task submissions, the system by Wetzel et al. (2015) is most similar to ours. On the source side (English in their case), they extract morphological information as well as coreference relations (they use Stanford CoreNLP (Lee et al., 2013) whereas we apply Bart 2 toolkit (Uryupina et al., 2012) for this task), and they detect the anaphoricity of the *it* pronoun using the NADA tool (Bergsma and Yarowsky, 2011). Another common feature is that both systems take advantage of the target language model. Wetzel et al. (2015)'s maximum entropy classifier Mallet (McCallum, 2002) uses the same logistic loss function as we do with the Vowpal Wabbit tool but the training data handling is different in these two tools. Mallet is a batch learner, optimizing over the whole data in a single step while Vowpal Wabbit optimizes incrementally after every example.

On the other hand, unlike us, Wetzel et al. (2015) do not use any syntactic information. The only syntax-based system in the DiscoMT 2015 shared task is the system of Loáiciga (2015). They make use of the Fips rule-based phrase-structure parser (Wehrli, 2007) whereas we acquire dependencies and syntactic functions using the MST parser (McDonald et al., 2005) and the MATE tools (Seeker and Kuhn, 2012) on the source side for English and German, respectively.

8 Conclusion

We presented our system submitted to the WMT16 shared task on cross-lingual pronoun prediction. It is based on Vowpal Wabbit and uses features from three sources: first, target language model (which served as the baseline in the shared task), second, the automatic linguistic analysis of the source text up to the levels of syntax and coreference, and third, a basic morphological analysis of the target text. Our systems were able to improve on the baseline in both language directions, with source language and target language model features having the largest impact on the results. Finally, we employ instance weighing, which proved to be a successful way to compensate for the differences between learning loss function and the official evaluation measure and to improve recall in infrequent pronoun classes.

Acknowledgments

This work has been supported by GAUK grant 338915 of the Charles University, Czech Science Foundation grant GA-16-05394S, the 7th Framework Programme of the EU grant QTLeap (No. 610516), and SVV project number 260 333. This work 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

Shane Bergsma and Dekang Lin. 2006. Bootstrapping Path-Based Pronoun Resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

Shane Bergsma and David Yarowsky. 2011. 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. Springer-Verlag.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A High-performance Syntactic and Semantic Dependency Parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A Transition-based System for Joint Part-of-speech Tagging and Labeled Non-projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*. Springer Verlag. In press.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany. Association for Computational Linguistics.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal. Association for Computational Linguistics.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Sweden.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Los Alamitos, CA, USA. IEEE Computer Society Press.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916.

Sharid Loáiciga. 2015. Predicting Pronoun Translation Using Syntactic, Morphological and Contextual Features from Parallel Data. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 78–85, Lisbon, Portugal, September. Association for Computational Linguistics.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. 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. Association for Computational Linguistics.

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

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey. European Language Resources Association (ELRA).

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. 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. Association for Computational Linguistics.

Olga Uryupina, Alessandro Moschitti, and Massimo Poesio. 2012. BART Goes Multilingual: The UniTN/Essex Submission to the CoNLL-2012 Shared Task. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 122–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eric Wehrli. 2007. Fips, a "Deep" Linguistic Multilingual Parser. In *Proceedings of the Workshop on Deep Linguistic Processing*, pages 120–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dominikus Wetzel, Adam Lopez, and Bonnie Webber. 2015. A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 115–121, Lisbon, Portugal. Association for Computational Linguistics.