Exploring Confidence-based Self-training for Multilingual Dependency Parsing in an Under-Resourced Language Scenario

Juntao Yu
University of Birmingham
Birmingham, UK
j.yu.1@cs.bham.ac.uk

Bernd Bohnet
Google
London, UK
bohnetbd@google.com

Abstract

This paper presents a novel self-training approach that we use to explore a scenario which is typical for under-resourced languages. We apply self-training on small multilingual dependency corpora of nine languages. Our approach employs a confidence-based method to gain additional training data from large unlabeled datasets. The method has been shown effective for five languages out of the nine languages of the SPMRL Shared Task 2014 datasets. We obtained the largest absolute improvement of two percentage points on Korean data. Our self-training experiments show improvements upon the best state-of-the-art systems of the SPMRL shared task that employs one parser only.

1 Introduction

The availability of the manually annotated treebanks and state-of-the-art dependency parsers (McDonald and Pereira, 2006; Nivre, 2009; Martins et al., 2010; Goldberg and Elhadad, 2010; Zhang and Nivre, 2011; Bohnet et al., 2013) leads to high accuracy on some languages such as English (Marcus et al., 1994), German (Kübler et al., 2006) and Chinese (Levy and Manning, 2003) that have large manually annotated datasets.

In contrast to resource-rich languages, languages that have less training data show a lower accuracy (Buchholz and Marsi, 2006; Nivre et al., 2007; Seddah et al., 2013; Seddah et al., 2014). Semi-supervised techniques gain popularity as they are able to improve parsing accuracy by exploiting unlabeled data which avoids the cost of labeling new data.

Self-training is one of these appealing techniques that have been successfully used for instance in constituency parsing for English texts (McClosky et al., 2006a; McClosky et al., 2006b; Reichart and Rappoport, 2007; Sagae, 2010) while for dependency parsing this approach was only effective in a few cases, in contrast to co-training which works for dependency parsing well too. In a co-training approach, at least another parser is employed to label additional training data.

McClosky et al. (2006a) used self-training for English constituency parsing. In their approaches, self-training was most effective when the parser is retrained on the combination of the initial training set and the large unlabeled dataset generated by both the generative parser and reranker. This leads to many subsequent applications on English texts via self-training for constituency parsing, cf. (McClosky et al., 2006b; Reichart and Rappoport, 2007; Sagae, 2010; Petrov and McDonald, 2012).

In contrast to English constituency parsing, self-training usually has proved to be less effective or has even shown negative results when applied to dependency parsing, cf. (Kawahara and Uchimoto, 2008; Plank, 2011; Cerisara, 2014; Björkelund et al., 2014). This paper makes the following contributions:

1. We present an effective confidence-based self-training approach.

2. We evaluate our approach on nine languages in a resource-poor parsing scenario.

3. We successfully improved the parsing performances on five languages which are Basque, German, Hungarian, Korean and Swedish.

The remainder of this paper is structured as follows: In Section 2, we discuss related work. In Section 3, we introduce our confidence-based approach to self-training and Section 4 describes the experimental set-up. Section 5 presents the results and contains a discussion of the results. Section 6 presents our conclusions.

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 350–358,
Uppsala, Sweden, August 24–26 2015.

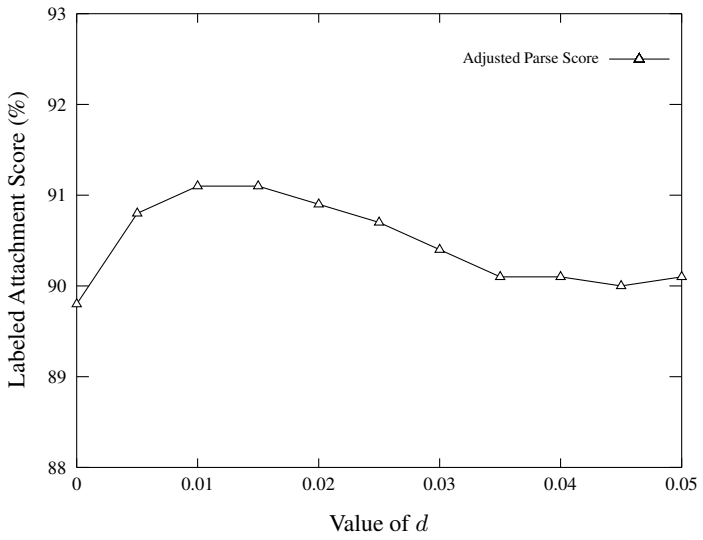

Figure 1: Accuracies of sentences which have a position number within the top 50% after ranking the auto-parsed sentences of development set by the adjusted parse scores with different values of d.

2 Related Work

Most of the reported positive results of self-training are evaluated on constituency parsing of English texts. McClosky et al. (2006a) reported strong results with an improvement of 1.1 F-score using the Charniak-parser, cf. (Charniak and Johnson, 2005). McClosky et al. (2006b) applied the method later on English out-of-domain texts which show good accuracy gains too.

Reichart and Rappoport (2007) showed that self-training can improve the performance of a constituency parser without a reranker when a small training set is used.

Sagae (2010) investigated the contribution of the reranker for a constituency parser. The results suggest that constituency parsers without a reranker can achieve significant improvements, but the results are still higher when a reranker is used.

In the SANCL 2012 shared task self-training was used by most of the constituency-based systems, cf. (Petrov and McDonald, 2012), which includes the top ranked system, this indicates that self-training is already an established technique to improve the accuracy of constituency parsing on English out-of-domain data, cf. (Le Roux et al., 2012). However, none of the dependency-based systems used self-training in the SANCL 2012 shared task.

One of the few successful approaches to self-training for dependency parsing was introduced by Chen et al. (2008). Chen et al. (2008) improved the unlabeled attachment score about one percentage point for Chinese. Chen et al. (2008) added sub-trees that span only over a few words, which means they have only short dependency edges. It is known that dependencies of short length have a higher accuracy than longer ones, cf. (McDonald and Nivre, 2007). Kawahara and Uchimoto (2008) used a separately trained binary classifier to select sentences as additional training data. Their approach improved the unlabeled accuracy of English texts in Chemical domain by about 0.5%.

Plank (2011) applied self-training with single and multiple iterations for parsing of Dutch using the Alpino parser (Malouf and Noord, 2004), which was modified to produce dependency trees. She found self-training produces only a slight improvement in some cases but worsened when more unlabeled data was added.

Cerisara (2014) and Björkelund et al. (2014) applied self-training to dependency parsing on nine languages. Cerisara (2014) found negative impacts only when they apply a basic self-training approach to a dependency parser. Similarly, Björkelund et al. (2014) observed a positive effect on Swedish only.

Recently, Dredze et al. (2008) and Crammer et al. (2009) introduced **confidence-based** learning methods that are able to measure the prediction quality. Their technique has been applied for a sequence labeling and a dependency parser which both use online-learning algorithms, cf. (Mejer

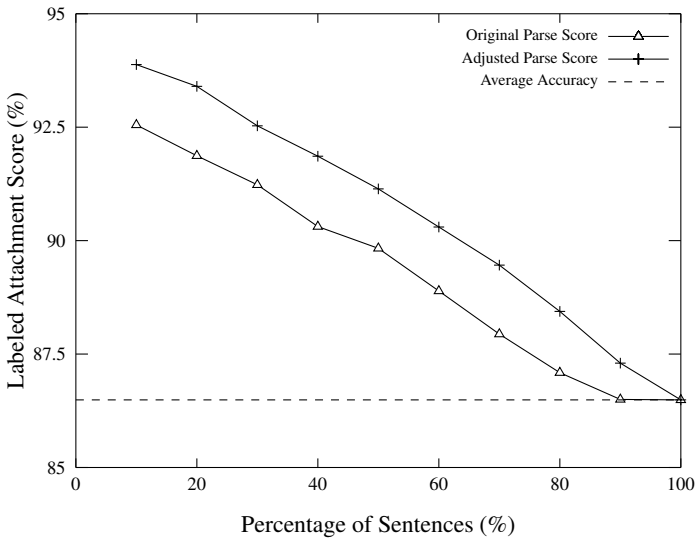

Figure 2: The accuracies when inspecting 10-100% sentences of the development set ranked by the confidence-based methods.

and Crammer, 2010; Mejer and Crammer, 2012). They evaluated several confidence-based methods and the empirical results showed that the confidence scores generated by some methods are highly relevant to the prediction accuracy, i.e. higher confidence is correlated with high accuracy scores.

The work most close to our approach is introduced by Goutam and Ambati (2011), who applied a multi-iteration self-training approach to improving Hindi in-domain parsing. In each iteration, they add 1,000 additional sentences to a small initial training set (2,972 sentences), the additional sentences are selected due to their parse scores. They improved the baseline by up to 0.7% and 0.4% for labeled and unlabeled attachment scores after 23 self-training iterations.

Our approach differs in three aspects from that of Goutam and Ambati (2011): We employ a single iteration self-training rather than multiple iterations. We add larger amounts of additional parsed unlabeled sentences to the initial training set for retraining and we applied our method in an under-resourced language scenario to nine languages.

3 Self-training

The hypotheses for our experiments is that the selection of high-quality dependency trees is a crucial precondition for the successful use of self-training in dependency parsing. Therefore, we explore a confidence-based method to select high-quality dependency trees from newly parsed sentences. Our self-training approach consists of a single iteration with the following steps:

1. A parser is trained on a (small) initial training set to generate a base model.
2. We analyze a large number of unlabeled sentences with the base model.
3. We build a new training set consisting of the initial training set and 50%[1] newly analyzed sentences parsed with a high confidence.
4. We retrain the parser on the new training set to produce a self-trained model.
5. Finally, the self-trained model is used to annotate the test set.

We use the freely available Mate tools[2] to implement the self-training approach. This tool set contains a part-of-speech (PoS) tagger, morphologic tagger, lemmatizer, graph-based parser and an arc-standard transition-based parser. The arc-standard transition-based parser has the option to use a graph-based model to rescore the beam which seems to be a sort-of reranking (Bohnet and Kuhn, 2012). The parser has further the option to use a joint tagging and parsing model with the

[1] We use 50% due to previous experiments on English that showed an optimal performance when adding 50% parsed sentences to the training set.

[2] https://code.google.com/p/mate-tools/

joint inference that improves both part-of-speech tagging and parsing accuracy.

We use the arc-standard transition-based parser employing beam search and a graph-based rescoring model. This parser computes a score for each dependency tree by summing up the scores for each transition and dividing the score by the total number of transitions, due to the swap-operation (used for non-projective parsing), the number of transition can vary, cf. (Kahane et al., 1998; Nivre, 2007).

For our self-training approach, we use the parse scores as confidence measure to select sentences. We observed that although the original parse score is the averaged value of a sequence of transitions of a parse, long sentences generally exhibit a higher score. Therefore, the score does not correlate well with the Labeled Attachment Score (LAS) as shown in Figure 2. Thus, we adjusted the score of the parser to maximize the correlation between the parse score and the labeled attachment score for each parse tree by subtracting the sentence length (L) multiplied by a fixed number d. The new parse scores are calculated as follow:

$$Score_{adjusted} = Score_{original} - L \times d \quad (1)$$

To obtain the constant d, we apply the defined formula with different values for d to all sentences of the development set and rank the sentences by their adjusted scores in a descending order. Let $No(i)$ be the position number of the i_{th} sentence after ranking them by the adjusted scores. The value of d is selected to maximize the accuracy of sentences that have a $No(i)$ within the top 50%. We evaluate stepwise different values of d from 0 to 0.05 with an increment of 0.005. The highest accuracy of the top ranked sentences is achieved when $d = 0.015$ (see Figure 1), thus d is set to 0.015 in our experiments. Figure 2 shows the accuracies when inspecting 10 -100% of sentences ranked by the adjusted and original parse scores. We found that the adjusted parse scores lead to a higher correlation with the accuracy of the parsed sentences compared to the original parse scores.

4 Experimental Set-up

We evaluate our approach on nine languages available from 2014 Shared Task at the Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL), cf. (Seddah et al., 2013; Seddah et al., 2014). We have chosen the datasets as they provide smaller data sets of 5k sentences for each language of the SPMRL shared task which are a good basis for our exploration for improving parsing accuracy of under-resourced languages and the shared task provides competitive results for these languages from the participants of the shared task that provides us strong accuracy scores against which we can compare our results.

Further, the organizers of the SPMRL shared task provided sufficient unlabeled data that are required for self-training. More precisely, for all language, we use as our initial training set the 5k datasets, we test on test sets available from the shared task and use a 100k SPMRL unlabeled data for each of the languages. We use the German development set (5,000 sentences) when tuning the fixed value d that was mentioned in Section 3. Table 1 shows statistics about the corpora that we use in our experiments.

As previously noted, the Mate transition-based dependency parser with default settings is used in our experiments, cf. (Bohnet et al., 2013). We use the parser's internal tagger to supply the part-of-speech for both unlabeled data and test data. The baselines are generated by training the parser on initial training data and testing the parser on the described test sets.

For the evaluation of the parser's accuracy, we report labeled attachment scores (LAS). In line with the SPMRL shared task evaluation, we include all punctuation marks in the evaluation.

For significance testing, we take Dan Bikel's randomized parsing evaluation comparator that was used by the CoNLL 2007 shared task with the default settings of 10,000 iterations (Nivre et al., 2007). The statistically significant results are marked due to their p-values (*) p-value$<$0.05, (**) p-value$<$0.01.

5 Results and Discussion

We evaluate our self-training approach on the test sets of nine languages. The unlabeled data was parsed and ranked by the confidence scores. Then we selected the 50k top ranked sentences and added those to the training sets.

The empirical results show that our approach worked for five languages which are Basque, German, Hungarian, Korean and Swedish. Our self-training method achieves the largest improvement on Korean with an absolute gain of 2.14 percent-

	Arabic	Basque	French	German	Hebrew
train:					
Sentences	5,000	5,000	5,000	5,000	5,000
Tokens	224,907	61,905	150,984	87,841	128,046
Avg. Length	44.98	12.38	30.19	17.56	25.60
test:					
Sentences	1,959	946	2,541	5,000	716
Tokens	73,878	11,457	75,216	92,004	16,998
Avg. Length	37.71	12.11	29.60	18.40	23.74
unlabeled:					
Sentences	100,000	100,000	100,000	100,000	100,000
Tokens	4,340,695	1,785,474	1,618,324	1,962,248	2,776,500
Avg. Length	43.41	17.85	16.18	19.62	27.77
	Hungarian	**Korean**	**Polish**	**Swedish**	
train:					
Sentences	5,000	5,000	5,000	5,000	
Tokens	109,987	68,336	52,123	76,357	
Avg. Length	21.99	13.66	10.42	15.27	
test:					
Sentences	1,009	2,287	822	666	
Tokens	19,908	33,766	8,545	10,690	
Avg. Length	19.73	14.76	10.39	16.05	
unlabeled:					
Sentences	100,000	100,000	100,000	100,000	
Tokens	1,913,154	2,147,605	2,024,323	1,575,868	
Avg. Length	19.13	21.48	20.24	15.76	

Table 1: Statistics about the corpora that we used in our experiments for the training set, test set and the unlabeled datasets for our multilingual evaluations, cf. (Seddah et al., 2014).

	Baseline	Self-train	LORIA
Arabic	82.09	**82.22**	81.65
Basque	78.35	79.22**	**81.39**
French	**81.91**	81.48	81.74
German	81.54	81.87**	**83.35**
Hebrew	78.86	**79.04**	75.55
Hungarian	83.13	**83.56***	82.88
Korean	73.31	**75.45****	74.15
Polish	**81.97**	81.35	79.95
Swedish	79.67	**80.26**	80.04
Average	80.09	**80.49**	80.08

Table 2: The table shows the results obtained for the languages of the SPMRL Shared Task 2014. The first column (Baseline) shows the results of our baseline parser (Mate), the second column shows the self-training experiments (Self-train) and the final column provides the results of the best non-ensemble system in the SPMRL Shared Task (LORIA).

age points. We also gain statistically significant improvements on Basque, German and Hungarian. Our self-training gains on these languages are 0.87%, 0.33% and 0.42% respectively.

We achieve an improvement of 0.59% on Swedish which is relatively high absolute improvement while it was not a statistically significant with a p-value of 0.067. To confirm the effectiveness of our method on Swedish, we further evaluate our method on the Swedish development set[3] (494 sentences).

Our self-training method achieves an accuracy of 76.16%*, which is 0.82 percentage points better than our baseline (75.34%). This improvement was statistically significant.

In terms of the effect of our method on other languages, our method gains moderate improvements on Arabic and Hebrew but these were not statistically significant accuracy gains. We found negative results for French and Polish. Table 2 shows a detailed evaluation of our self-training experiments.

We compare our self-training results with the best results of non- ensemble parsing system of SPMRL shared tasks (Seddah et al., 2013; Seddah et al., 2014). The average accuracy of our baseline on nine languages is same as the one achieved by the best single parser system of SPMRL 2014 shared task (Cerisara, 2014), their system employs LDA clusters (Chrupala, 2011) to exploit unlabeled data as well.

Our self-training results is on average 0.41% higher than those of Cerisara (2014). Our self-training method performs better on six languages (Arabic, Hebrew, Hungarian, Korean, Polish and Swedish) compared to the best non-ensemble system.

The confidence scores have shown to be crucial for the successful application of self-training for dependency parsing. In contrast to constituency parsing, self-training for dependency parsing does not work without this additional confidence-based selection step. The question about a possible reason for the different behavior of self-training in dependency parsing and in constituency parsing remains open and only speculative answers could be given. We plan to investigate this further in future.

Self-training behaves somewhat different from co-training in that co-training seems to be able to exploit the differences in the parse trees produced by two or more parsers. While self-training relies on a single parser due to its definition, co-training uses at least another parser what is the main difference to self-training. Co-training does not employ in its most simple form selection, but confidence helps in a co-training scenario too since selecting those dependency trees for retraining on which two or more parsers agree improves further the accuracy. Hence, confidence-based methods is a more effective for co-training, cf. (Blum and Mitchell, 1998; Sarkar, 2001; Steedman et al., 2003).

An open question remains why for some of the languages the approach did not work. In future work, we want to address this question. A first observation is that the quality of the unlabeled data

[3]We did not use the Swedish development set for tuning in our experiments.

might have an effect. For instance, the average length of unlabeled data of Polish and French is different from that of the training and test set for these languages.

6 Conclusions

In this paper, we present an effective confidence-based self-training approach for multilingual dependency parsing. We evaluated our approach on nine languages in a scenario for under-resourced languages when only a small amount of training data is available.

We apply the same setting for all language by retraining the parser on the new training set that consists of the initial training set and the top 50k ranking parse trees from the 100k parsed sentences of the unlabeled data.

As a result, our approach successfully improves the accuracies of five languages which are Basque, German, Hungarian, Korean and Swedish without tuning variables for individual language. We can report the largest accuracy gain of 2.14% on Korean, on average we improve the baselines of five languages by 0.87%. Previous work that apply self-training to dependency parsing showed often negative results (Plank, 2011; Cerisara, 2014) or was evaluated on one language only (Chen et al., 2008; Goutam and Ambati, 2011; Björkelund et al., 2014).

This is to the best of our knowledge the first time that self-training is found effective for a number of languages. In addition, our self-training results are better than the best reported results generated from a non-ensemble system that used LDA clusters, cf. Cerisara (2014).

Finally, our approach contributes a novel confidence-based self-training method that is able to access the parse quality of unlabeled data and to carry out a pre-selection of the parsed sentences. We conclude that self-training based on confidence is worth using in an under-resourced language scenario and that a confidence-based self-training approach seems to be crucial for the successful application of self-training in dependency parsing. This paper underlines the finding that the pre-selection of parsed dependency trees from unlabeled sources is probably a precondition for the effectivity of self-training and leads additionally to a higher accuracy gain.

References

Anders Björkelund, Özlem Çetinoğlu, Agnieszka Faleńska, Richárd Farkas, Thomas Mueller, Wolfgang Seeker, and Zsolt Szántó. 2014. The IMS-Wrocław-Szeged-CIS entry at the SPMRL 2014 Shared Task: Reranking and Morphosyntax meet Unlabeled Data. In *Proc. of the Shared Task on Statistical Parsing of Morphologically Rich Languages*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory (COLT)*, pages 92–100.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both worlds – a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chpater of the Association for Computational Linguistics (EACL)*, pages 77–87.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas Filip Ginter, and Jan Hajic. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Associtation for Computational Linguistics*, 1.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*, pages 149–164.

Christophe Cerisara. 2014. Semi-supervised experiments at LORIA for the SPMRL 2014 Shared Task. In *Proc. of the Shared Task on Statistical Parsing of Morphologically Rich Languages*, Dublin, Ireland, August.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wenliang Chen, Youzheng Wu, and Hitoshi Isahara. 2008. Learning reliable information for dependency parsing adaptation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 113–120. Association for Computational Linguistics.

Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372. Asian Federation of Natural Language Processing.

Koby Crammer, Alex Kulesza, and Mark Dredze. 2009. Adaptive regularization of weight vectors. In *Advances in Neural Information Processing Systems*, pages 414–422.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *Proceedings of the 25th international conference on Machine learning*, pages 264–271. ACM.

Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 742–750.

Rahul Goutam and Bharat Ram Ambati. 2011. Exploring self training for hindi dependency parsing. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, volume 2, pages 22–69.

Sylvain Kahane, Alexis Nasr, and Owen Rambow. 1998. Pseudo-projectivity: A polynomially parsable non-projective dependency grammar. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguistics (COLING)*, pages 646–652.

Daisuke Kawahara and Kiyotaka Uchimoto. 2008. Learning reliability of parses for domain adaptation of dependency parsing. In *IJCNLP*, volume 8.

Sandra Kübler, Erhard W. Hinrichs, and Wolfgang Maier. 2006. Is it really that difficult to parse German? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasul Samad Zadeh Kaljahi, and Anton Bryl. 2012. Dcu-paris13 systems for the sancl 2012 shared task.

Roger Levy and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 439–446.

Robert Malouf and Gertjan Noord. 2004. Wide coverage parsing with stochastic attribute value grammars. In *In Proc. of IJCNLP-04 Workshop Beyond Shallow Analyses*.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate-argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 114–119.

André FT Martins, Noah A Smith, Eric P Xing, Pedro MQ Aguiar, and Mário AT Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 34–44. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 337–344. Association for Computational Linguistics.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 122–131.

Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 81–88.

Avihai Mejer and Koby Crammer. 2010. Confidence in structured-prediction using confidence-weighted models. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 971–981. Association for Computational Linguistics.

Avihai Mejer and Koby Crammer. 2012. Are you sure?: Confidence in prediction of dependency tree edges. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 573–576, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre. 2007. Incremental non-projective dependency parsing. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 396–403.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 351–359. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59.

Barbara Plank. 2011. *Domain Adaptation for Parsing*. Ph.d. thesis, University of Groningen.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *ACL*, volume 7, pages 616–623.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 37–44. Association for Computational Linguistics.

Anoop Sarkar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 175–182.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland, August. Dublin City University.

Mark Steedman, Rebecca Hwa, Miles Osborne, and Anoop Sarkar. 2003. Corrected co-training for statistical parsers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 95–102.

Yue Zhang and Joakim Nivre. 2011. Transition-based parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.