Introducing the IMS-Wrocław-Szeged-CIS Entry at the SPMRL 2014 Shared Task: Reranking and Morphosyntax Meet Unlabeled Data*

Anders Björkelund[§] and **Özlem Çetinoğlu**[§] and **Agnieszka Faleńska**[◇,§]

Richárd Farkas[†] and **Thomas Müller**[‡] and **Wolfgang Seeker**[§] and **Zsolt Szántó**[†]

[§]Institute for Natural Language Processing University of Stuttgart, Germany
[◇]Institute of Computer Science, University of Wrocław, Poland
[†]Department of Informatics University of Szeged, Hungary
[‡]Center for Information and Language Processing University of Munich, Germany

{anders,ozlem,muellets,seeker}@ims.uni-stuttgart.de
agnieszka.falenska@cs.uni.wroc.pl
{rfarkas,szantozs}@inf.u-szeged.hu

Abstract

We summarize our approach taken in the SPMRL 2014 Shared Task on parsing morphologically rich languages. Our approach builds upon our contribution from last year, with a number of modifications and extensions. Though this paper summarizes our contribution, a more detailed description and evaluation will be presented in the accompanying volume containing notes from the SPMRL 2014 Shared Task.

1 Introduction

This paper summarizes the approach of IMS-Wrocław-Szeged-CIS taken for the SPMRL 2014 Shared Task on parsing morphologically rich languages (Seddah et al., 2014). Since this paper is a rough summary that is written before submission of test runs we refer the reader to the full description paper which will be published after the shared task (Björkelund et al., 2014).[1]

The SPMRL 2014 Shared Task is a direct extension of the SPMRL 2013 Shared Task (Seddah et al., 2013) which targeted parsing morphologically rich languages. The task involves parsing both dependency and phrase-structure representations of 9 languages: Arabic, Basque, French, German, Hebrew, Hungarian, Korean, Polish, and Swedish. The only difference between the two tasks is that large amounts of unlabeled data are additionally available to participants for the 2014 task.

Our contribution builds upon our system from last year (Björkelund et al., 2013), with additional features and components that try to exploit the unlabeled data. Given the limited window of time to participate in this year's shared task, we only contribute to the setting with predicted preprocessing, using the largest available training data set for each language.[2] We also do not participate in the Arabic track since the shared task organizers did not provide any unlabeled data at a reasonable time.

2 Review of Last Year's System

Our current system is based on the system we participated with in the SPMRL 2013 Shared Task. We summarize the architecture of this system as three different components.

*Authors in alphabetical order

[1]Due to logistical constraints this paper had to be written before the deadlines for the actual shared task and do thus not contain a full description of the system, nor the experimental evaluation of the same.

[2]In other words, no gold preprocessing or smaller training sets.

First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages, pages 97–102 Dublin, Ireland, August 23-29 2014.

2.1 Preprocessing

As the initial step of preprocessing we converted the Shared Task data from the CoNLL06 format to CoNLL09, which required a decision on using coarse or fine grained POS tags. After a set of preliminary experiments we picked fine POS tags where possible, except Basque and Korean.

We used MarMoT[3] (Müller et al., 2013) to predict POS tags and morphological features jointly. We integrated the output from external morphological analyzers as features to MarMoT. We also experimented with the integration of predicted tags provided by the organizers and observed that these *stacked* models help improve Basque, Polish, and Swedish preprocessing. The stacked models provided additional information to our tagger since the provided predictions were coming from models trained on larger training sets than the shared task training sets.

2.2 Dependency Parsing

The dependency parsing architecture of our SPMRL 2013 Shared Task contribution is summarized in Figure 1. The first step combines the n-best trees of two parsers, namely the mate parser[4] (Bohnet, 2010) and a variant of the EasyFirst parser (Goldberg and Elhadad, 2010), which we call best-first parser. We merged the 50-best analyses from these parsers into one n-best list of 50 to 100 trees. We then added parsing scores to the n-best trees from the two parsers, and additionally from the turboparser[5] (Martins et al., 2010).

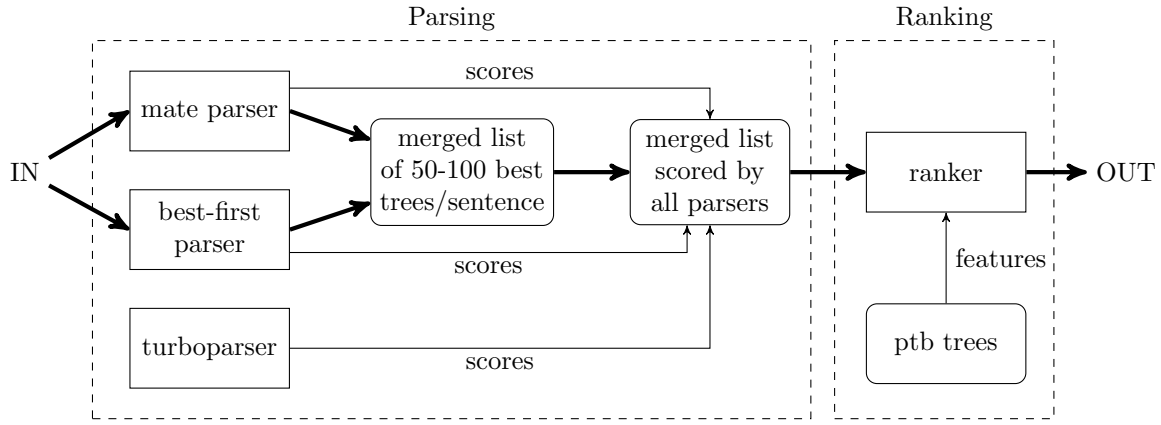

Figure 1: Architecture of the dependency ranking system from (Björkelund et al., 2013).

The scored trees are fed into the ranking system. The ranker utilizes the parsing scores and features coming from both constituency and dependency parses. We specified a default feature set and experimented with additional features for each language for optimal results. We achieved over 1% LAS improvement on all languages except a 0.3% improvement on Hungarian.

2.3 Constituency Parsing

The constituency parsing architecture advances in three steps. For all setups we removed the morphological annotation of POS tags and the function labels of non-terminals and apply the Berkeley Parser (Petrov et al., 2006) as our baseline. As the first setup, we replaced words with a frequency < 20 with their predicted part-of-speech and morphology tags and improved the PARSEVAL scores across languages. The second setup employed a product grammar (Petrov, 2010), where we combined 8 different grammars trained on the same data but with different initialization setups. As a result, the scores substantially improved on all languages.

Finally, we conducted ranking experiments on the 50-best outputs of the product grammars. We used a slightly modified version of the Mallet toolkit (McCallum, 2002), where the reranker is trained for the

[3] https://code.google.com/p/cistern/
[4] https://code.google.com/p/mate-tools
[5] http://www.ark.cs.cmu.edu/TurboParser/

maximum entropy objective function of Charniak and Johnson (2005) and uses the standard feature set from Charniak and Johnson (2005) and Collins (2000). Hebrew and Polish scores remained almost the same, whereas Basque, French, and Hungarian highly benefited from reranking.

3 Planned Additions to Last Year's System

This year we extend our systems for both the constituency and dependency tracks to add additional information and try to profit from unlabeled data.

3.1 Preprocessing

We use the mate-tools' lemmatizer and MarMoT to preprocess all labeled and unlabeled data. From the SPMRL 2013 Shared Task, we learned that getting as good preprocessing as possible is an important part of the overall improvements. Preprocessing consists of predicting lemmas, part-of-speech, and morphological features. Preprocessing for the training data is done via 5-fold jackknifing to produce realistic input features for the parsers. This year we do not do stacking on top of provided morphological analyses since the annotations on the labeled and unlabeled data were inconsistent for some languages.[6]

3.2 Dependency Parsing

We pursue two different ways of integrating additional information into our system from the SPMRL 2013 Shared Task (Björkelund et al., 2013): **supertags** and **co-training**.

Supertags (Bangalore and Joshi, 1999) are tags that encode more syntactic information than standard part-of-speech tags. Supertags have been used in deep grammar formalisms like CCG or HPSG to prune the search space for the parser. The idea has been applied to dependency parsing by Foth et al. (2006) and recently to statistical dependency parsing (Ouchi et al., 2014; Ambati et al., 2014), where supertags are used as features rather than to prune the search space. Since the supertag set is dynamically derived from the gold-standard syntactic structures, we can encode different kinds of information into a supertag, in particular also morphological information. Supertags are predicted before parsing using MarMoT and are then used as features in the mate parser and the turboparser.

We will use a variant of **co-training** (Blum and Mitchell, 1998) by applying two different parsers to select additional training material from unlabeled data. We use the mate parser and the turboparser to parse the unlabeled data provided by the organizers. We then select sentences where both parsers agree on the structure as additional training examples following Sagae and Tsujii (2007). We then train two more models: one on the labeled training data and the unlabeled data selected by the two parsers, and one only on the unlabeled data. These two models are then integrated into our parsing system from 2013 as additional scorers to score the n-best list. Their scores are used as features in the ranker.

Before we parse the unlabeled data to obtain the training sentences, we filter it in order to arrive at a cleaner corpus. Most importantly, we only keep sentences up to length 50, and which contain at maximum two unknown words (compared to the labeled training data).

3.3 Constituency Parsing

We experiment with two approaches for improving constituency parsing:

Preterminal labelsets play an important role in constituency parsing of morphologically rich languages (Dehdari et al., 2011). Instead of removing the morphological annotation of POS tags, we use a preterminal set which carries more linguistic information while still keeping it compact. We follow the merge procedure for morphological feature values of Szántó and Farkas (2014). This procedure outputs a clustering of full morphological descriptions and we use the cluster IDs as preterminal labels for training the Berkeley Parser.

Reranking at the constituency parsing side is enriched by novel features. We define feature templates exploiting co-occurrence statistics from the unlabeled datasets; automatic dependency parses of the sentence in question (Farkas and Bohnet, 2012); Brown clusters (Brown et al., 1992); and atomic morphological feature values (Szántó and Farkas, 2014).

[6]The organizers later resolved this issue by patching the data, although time constraints prevented us from using the patched data.

4 Conclusion

This paper describes our plans for the SPMRL 2014 Shared Task, most of which are yet to be implemented. For the actual system description and our results, we refer the interested reader to (Björkelund et al., 2014) and (Seddah et al., 2014).

Acknowledgements

Agnieszka Faleńska is funded through the Project International computer science and applied mathematics for business study programme at the University of Wrocław co-financed with European Union funds within the European Social Fund No. POKL.04.01.01-00-005/13. Richárd Farkas and Zsolt Szántó are funded by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013). Thomas Müller is supported by a Google Europe Fellowship in Natural Language Processing. The remaining authors are funded by the Deutsche Forschungsgemeinschaft (DFG) via the SFB 732, projects D2 and D8 (PI: Jonas Kuhn).

We also express our gratitude to the treebank providers for each language: Arabic (Maamouri et al., 2004; Habash and Roth, 2009; Habash et al., 2009; Green and Manning, 2010), Basque (Aduriz et al., 2003), French (Abeillé et al., 2003), Hebrew (Sima'an et al., 2001; Tsarfaty, 2010; Goldberg, 2011; Tsarfaty, 2013), German (Brants et al., 2002; Seeker and Kuhn, 2012), Hungarian (Csendes et al., 2005; Vincze et al., 2010), Korean (Choi et al., 1994; Choi, 2013), Polish (Świdziński and Woliński, 2010), and Swedish (Nivre et al., 2006).

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In Anne Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *TLT-03*, pages 201–204.

Bharat Ram Ambati, Tejaswini Deoskar, and Mark Steedman. 2014. Improving dependency parsers using combinatory categorial grammar. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 159–163, Gothenburg, Sweden, April. Association for Computational Linguistics.

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Anders Björkelund, Özlem Çetinoğlu, Richárd Farkas, Thomas Müller, and Wolfgang Seeker. 2013. (re)ranking meets morphosyntax: State-of-the-art results from the SPMRL 2013 shared task. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 135–145, Seattle, Washington, USA, October. Association for Computational Linguistics.

Anders Björkelund, Özlem Çetinoğlu, Agnieszka Faleńska, Richárd Farkas, Thomas Müller, 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 *Notes of the SPMRL 2014 Shared Task on Parsing Morphologically-Rich Languages*, Dublin, Ireland, August.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41, Sozopol, Bulgaria.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. Kaist tree bank project for korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14. Citeseer.

Jinho D. Choi. 2013. Preparing korean data for the shared task on parsing morphologically rich languages. *CoRR*, abs/1309.1649.

Michael Collins. 2000. Discriminative Reranking for Natural Language Parsing. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 175–182.

Dóra Csendes, Janós Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Text, Speech and Dialogue: Proceedings of TSD 2005*. Springer.

Jon Dehdari, Lamia Tounsi, and Josef van Genabith. 2011. Morphological features for parsing morphologically-rich languages: A case of arabic. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 12–21, Dublin, Ireland, October. Association for Computational Linguistics.

Richárd Farkas and Bernd Bohnet. 2012. Stacking of dependency and phrase structure parsers. In *Proceedings of COLING 2012*, pages 849–866, Mumbai, India, December. The COLING 2012 Organizing Committee.

Kilian A. Foth, Tomas By, and Wolfgang Menzel. 2006. Guiding a constraint dependency parser with supertags. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 289–296, Sydney, Australia, July. Association for Computational Linguistics.

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*, pages 742–750, Los Angeles, California, June. Association for Computational Linguistics.

Yoav Goldberg. 2011. *Automatic syntactic processing of Modern Hebrew*. Ph.D. thesis, Ben Gurion University of the Negev.

Spence Green and Christopher D. Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 394–402, Beijing, China, August. Coling 2010 Organizing Committee.

Nizar Habash and Ryan Roth. 2009. Catib: The columbia arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore, August. Association for Computational Linguistics.

Nizar Habash, Reem Faraj, and Ryan Roth. 2009. Syntactic Annotation in the Columbia Arabic Treebank. In *Proceedings of MEDAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*.

Andre Martins, Noah Smith, Eric Xing, Pedro Aguiar, and Mario 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, Cambridge, MA, October. Association for Computational Linguistics.

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

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *In Proceedings of EMNLP*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of LREC*, pages 1392–1395, Genoa, Italy.

Hiroki Ouchi, Kevin Duh, and Yuji Matsumoto. 2014. Improving dependency parsers with supertags. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 154–158, Gothenburg, Sweden, April. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics.

Slav Petrov. 2010. Products of Random Latent Variable Grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Los Angeles, California, June. Association for Computational Linguistics.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1044–1050, Prague, Czech Republic, June. Association for Computational Linguistics.

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, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho Choi, Matthieu Constant, Richárd Farkas, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2014. Overview of the SPMRL 2014 shared task on parsing morphologically rich languages. In *Notes of the SPMRL 2014 Shared Task on Parsing Morphologically-Rich Languages*, Dublin, Ireland.

Wolfgang Seeker and Jonas Kuhn. 2012. Making Ellipses Explicit in Dependency Conversion for a German Treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3132–3139, Istanbul, Turkey. European Language Resources Association (ELRA).

Khalil Sima'an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a Tree-Bank for Modern Hebrew Text. In *Traitement Automatique des Langues*.

Marek Świdziński and Marcin Woliński. 2010. Towards a bank of constituent parse trees for Polish. In *Text, Speech and Dialogue: 13th International Conference (TSD)*, Lecture Notes in Artificial Intelligence, pages 197—204, Brno, Czech Republic. Springer.

Zsolt Szántó and Richárd Farkas. 2014. Special techniques for constituent parsing of morphologically rich languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 135–144, Gothenburg, Sweden, April. Association for Computational Linguistics.

Reut Tsarfaty. 2010. *Relational-Realizational Parsing*. Ph.D. thesis, University of Amsterdam.

Reut Tsarfaty. 2013. *A Unified Morpho-Syntactic Scheme of Stanford Dependencies*. Proceedings of ACL.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *LREC*.