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

Anders Björkelund[§] Özlem Çetinoğlu[§] Agnieszka Faleńska[◇,§] Richárd Farkas[†] Thomas Müller[‡] Wolfgang Seeker[§] Zsolt Szántó[†]

[§]IMS
University of Stuttgart, Germany
{anders, ozlem, muellets, seeker}@ims.uni-stuttgart.de

[‡]CIS
University of Munich, Germany

[◇]Institute of Computer Science
University of Wrocław, Poland
agnieszka.falenska@cs.uni.wroc.pl

[†]Department of Informatics
University of Szeged, Hungary
{rfarkas, szantozs}@inf.u-szeged.hu

Summary and Findings

- Best Scores in Constituency track (except Polish)
  - Unlabeled data helped alleviate lexical sparsity
  - But not as much in overall as replacing rare words with morphology predictions
  - Brown clusters and atomic morphological feature values helped in reranking
- Best Scores in Dependency track
  - Supertags helped regardless of their model
  - Blending mate+TurboParser+BestFirst constituted a strong baseline (ranked 2nd)
  - Experiments with unlabeled data resulted in negligible improvements (except Swedish)

Preprocessing

- Predicted POS and morphology using MarMoT (Müller et al., 2013)
- Extended MarMoT with features from morphological analyzers
  - Input to the analyzers is training, development, and unlabeled data
- Also utilized the predicted tags provided by the organizers
  - Assigning multiple predictions instead of the best prediction (i.e., stacking)

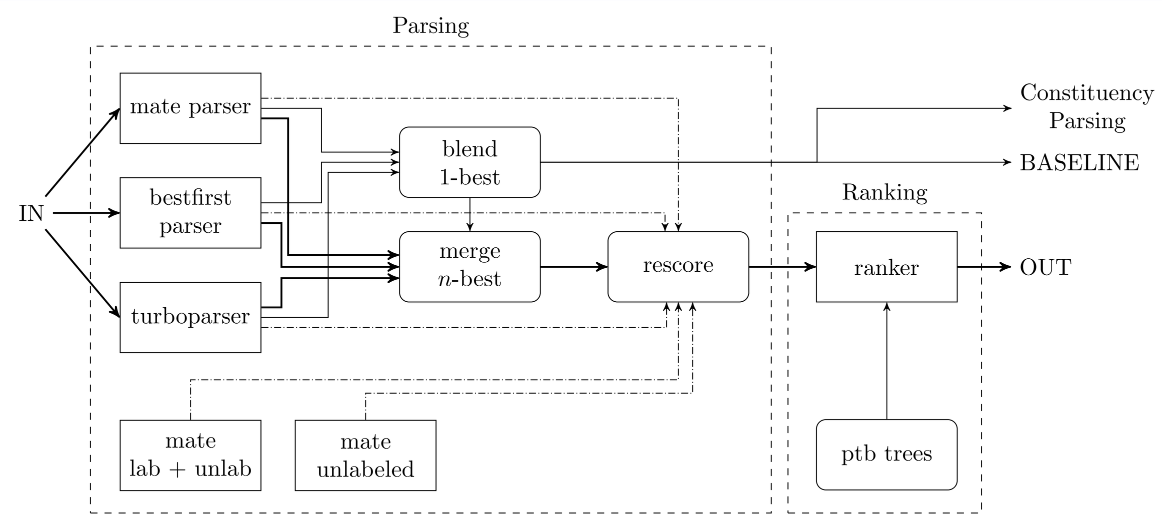
	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish
2013	*98.23/89.05*	97.61/90.92	98.10/91.80	97.09/97.67	98.72/97.59	94.03/87.68	*98.56/92.63*	*97.83/97.62*
2014	97.52/87.81	97.08/89.36	97.98/90.38	96.97/97.15	98.49/97.45	93.82/87.44	98.39/91.00	97.40/97.16

Constituency Parsing

- Baseline: The Berkeley parser (Petrov et al., 2006)
- Replaced rare words with their morphological tag from MarMoT (*Replace*)
- Utilized unlabeled data via external lexicons (Goldberg and Elhadad, 2013) (*ExtendLex*)
- Employed product grammars (Petrov, 2010) and reranked their output (Charniak and Johnson, 2005)
- Added new reranker features
  - *ExtendLex*: atomic morphological values, Brown clusters (Brown et al., 1992), dependency parsing features ⇒ up to 3.1% improvement (Basque)
  - *Replace*: dependency parsing features ⇒ up to 1.5% improvement (French)

	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish
$\text{Berkeley}_{mainPOS}$	72.32	79.35	82.26	88.71	83.84	71.85	86.75	75.19
$\text{Berkeley}_{fullMorph}$	77.82	79.17	80.22	88.40	87.18	82.28	85.06	72.82
ExtendLex	77.51	79.67	81.54	89.33	88.99	-	88.21	74.57
Replace	84.27	80.26	82.99	89.73	89.59	83.07	90.29	77.08
ExtendLex Product	80.71	**81.38**	82.13	**89.92**	90.43	-	91.52	78.21
Replace Product	**85.31**	81.29	**84.55**	89.87	**90.72**	**83.86**	**92.28**	**78.66**
$\text{ExtendLex Reranked}_{dflt}$	81.59	81.92	82.83	90.16	91.06	-	89.79	79.09
$\text{Replace Reranked}_{dflt}$	86.11	82.30	84.59	90.02	91.09	83.50	88.31	78.87
$\Delta_{ExtendLex\ Product}$	3.12	1.38	2.56	0.84	1.62	-	-0.08	0.57
$\text{ExtendLex Reranked}_{dflt+morph+Brown+dep}$	83.83	82.76	84.69	**90.76**	**92.05**	-	**91.44**	78.78
$\Delta_{Replace\ Product}$	1.42	1.49	1.5	0.6	1.17	0.92	-1.75	0.72
$\text{Replace Reranked}_{dflt+dep}$	**86.73**	**82.78**	**86.05**	90.47	91.89	**84.78**	90.53	**79.38**

Dependency Parsing



- Parsers
  - The mate parser (Bohnet, 2010) + Self-trained parsers for tree rescoring
  - TurboParser (Martins et al., 2010)
  - In-house BestFirst parser (Goldberg and Elhadad, 2010)
- Extended mate and TurboParser with supertags (Ouchi et al., 2014)
  ⇒ Boosts performance by up to 1.8% LAS (Polish)
- Baseline: Blending of the three base parsers (Sagae and Lavie, 2006)

	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish
mate	83.96	84.34	91.25	79.66	84.15	85.49	85.96	76.50
turbo	83.98	84.03	91.32	78.99	82.50	86.08	85.27	75.62
mate_{stag}	84.74	84.78	91.49	79.66	**84.47**	86.52	86.23	77.25
bestfirst	75.76	83.33	90.91	78.60	75.52	83.75	82.52	75.78
turbo_{stag}	**85.08**	84.47	91.69	80.05	83.39	**86.92**	**87.03**	77.18
blend	84.71	**85.10**	**92.19**	**80.65**	84.24	86.83	86.97	**78.23**
mate_{ulbl}	83.82	82.43	88.35	78.12	82.26	85.73	85.92	75.48
$\text{mate}_{lbl+ulbl}$	85.02	84.60	91.34	79.95	84.38	86.33	86.63	78.10
Δ_{best}	1.38	0.91	0.56	1.28	0.61	0.93	1.04	1.41
Δ_{blend}	1.75	0.91	0.56	1.28	0.84	1.02	1.10	1.41
Ranked	**86.46**	**86.01**	**92.75**	**81.93**	**85.08**	**87.85**	**88.07**	**79.64**
Oracle	*91.66*	*90.31*	*97.15*	*87.07*	*88.37*	*94.72*	*95.30*	*85.40*

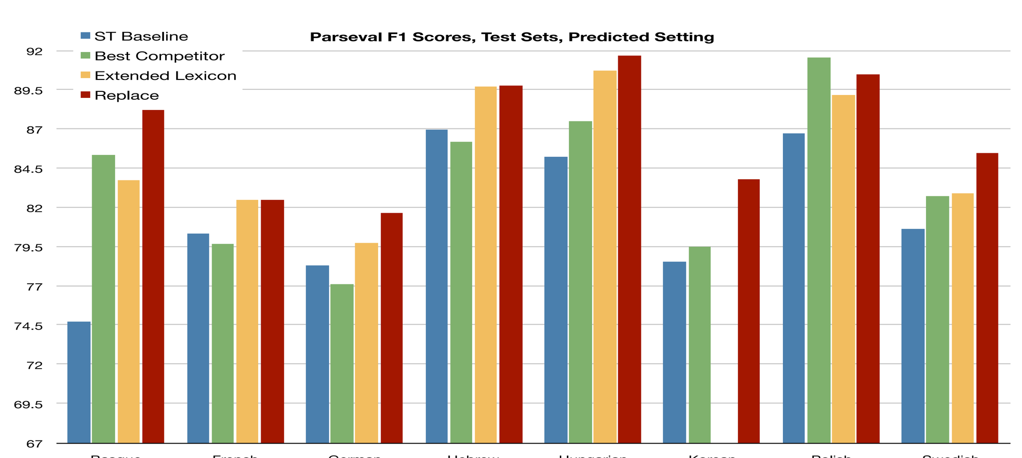
- Ranker features (tuned for each language)
  - Scores from base parsers (and combinations)
  - Projectivity features and ill-nestedness
  - Function label uniqueness for certain labels
  - Constituency features based on paths in constituency trees

Test Set Results

Constituency Results:

- Achieved the best scores on all languages except Polish
- *Replace* outperformed *ExtendLex* ⇒ by up to 4.5% (Basque)
- For 5 out of 7 languages we submitted both systems, our contributions came first and second

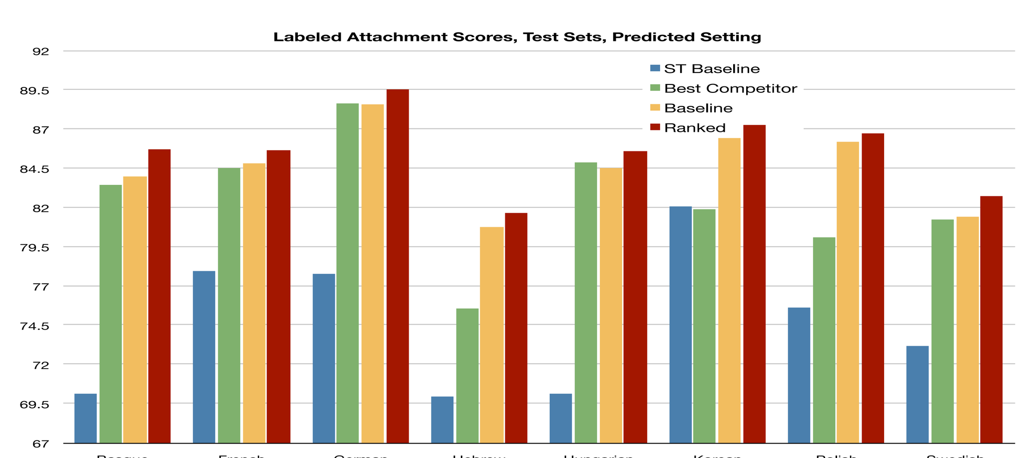
	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish
ST Baseline	74.74	80.38	78.30	86.96	85.22	78.56	86.75	80.64
Best Competitor	85.35	79.68	77.15	86.19	87.51	79.50	**91.60**	82.72
ExtendLex Reranked	83.78	**82.53**	79.76	89.75	90.76	-	89.19	82.94
Replace Reranked	**88.24**	82.52	**81.66**	**89.80**	**91.72**	**83.81**	90.50	**85.50**



Dependency Results:

- Achieved the best scores on all languages
- Our baseline came third for Hungarian and second for all other languages
- Ranking consistently improves over our baseline on all languages
  ⇒ up to 1.7% LAS improvement (Basque)

	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish
ST Baseline	70.11	77.98	77.81	69.97	70.15	82.06	75.63	73.21
Best Competitor	83.46	84.51	88.66	75.55	84.90	81.88	80.13	81.23
Baseline	83.97	84.83	88.62	80.77	84.51	86.42	86.21	81.42
Ranked	**85.70**	**85.66**	**89.58**	**81.65**	**85.59**	**87.27**	**86.75**	**82.75**



References

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *COLING*.
Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4).
Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *ACL*.
Goldberg, Y. and Elhadad, M. (2010). An Efficient Algorithm for Easy-First Non-Directional Dependency Parsing. In *NAACL-HLT*.
Goldberg, Y. and Elhadad, M. (2013). Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System. *Computational Linguistics*, 39(1).
Martins, A., Smith, N., Xing, E., Aguiar, P., and Figueiredo, M. (2010). Turbo Parsers: Dependency Parsing by Approximate Variational Inference. In *EMNLP*.
Müller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *EMNLP*.
Ouchi, H., Duh, K., and Matsumoto, Y. (2014). Improving dependency parsers with supertags. In *EACL*.
Petrov, S. (2010). Products of Random Latent Variable Grammars. In *NAACL-HLT*.
Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL*.
Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *NAACL-HLT*.

Acknowledgments

Agnieszka Faleńska is funded by the Project International computer science and applied mathematics for business study program at the University of Wrocław co-financed with EU funds within the European Social Fund (POKL.04.01.01-00-005/13). Richárd Farkas and Zsolt Szántó are funded by the EU and the European Social Fund through the project FuturICT.hu (TÁMOP-4.2.2.C-11/1/KONV-2012-0013). Thomas Müller is supported by a Google Europe Fellowship in NLP. The remaining authors are funded by the DFG via the SFB 732, projects D2 and D8 (PI: Jonas Kuhn).