Reranking and Morphosyntax Meet Unlabeled Data

Anders Björkelund, Özlem Çetinoğlu, Agnieszka Faleńska,
Richárd Farkas, Thomas Müller, Wolfgang Seeker, Zsolt Szántó

IMS-Wrocław-Szeged-CIS

Preprocessing

Constituency Parsing

Dependency Parsing

Conclusion

Our Contribution

- Built upon our system from last year [Björkelund et al., 2013]
- Participated only to the predicted setting using full training data
- All languages but Arabic

Section 1

Preprocessing

Preprocessing

Our system from last year

- Use of our own POS tags and morphological features
- MarMoT [Müller et al., 2013] to predict POS and morphological features jointly
- Two additional sources of features for MarMoT:
  - morphological analyzers
  - predicted tags provided by the organizers $\rightarrow$ stacking

Preprocessing

New this year

- Use of our own POS tags and morphological features
- MarMoT [Müller et al., 2013] to predict POS and morphological features jointly
- Two additional sources of features for MarMoT:
  - morphological analyzers also on unlabeled data
  - predicted tags provided by the organizers $\rightarrow$ predicted dictionaries

Predicted Dictionaries

- Listed all the predictions of a word present in the training, development and unlabeled data
- Assigned multiple predictions instead of the best prediction (i.e., stacking)

Preprocessing Accuracies

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
POS accuracy								
2013	*98.23*	97.61	98.10	97.09	98.72	94.03	*98.56*	*97.83*
Δ	-0.71	-0.53	-0.12	-0.12	-0.23	-0.21	-0.17	-0.43
2014	97.52	97.08	97.98	96.97	98.49	93.82	98.39	97.40
Morphological feature accuracy								
2013	*89.05*	90.92	91.80	97.67	97.59	87.68	*92.63*	*97.62*
Δ	-1.24	-1.56	-1.42	-0.52	-0.14	-0.24	-1.63	-0.46
2014	87.81	89.36	90.38	97.15	97.45	87.44	91.00	97.16

The italic figures denote languages where stacking was used last year (Basque, Polish, Swedish)

Section 2

Constituency Parsing

Constituency Parsing

Followed our pipeline from last year

- PCF Grammars with Latent Annotations [Petrov et al., 2006]
  Lexical Sparsity:
  - replace rare words by their morphological analysis
- Product Grammars [Petrov, 2010]
- Reranking
  - features of Charniak and Johnson [2005] and Collins [2000]

Constituency Parsing

New this year

- PCF Grammars with Latent Annotations [Petrov et al., 2006]
  Lexical Sparsity:
  - replace rare words by their morphological analysis
  - extended lexicon model [Goldberg and Elhadad, 2013]
- Product Grammars [Petrov, 2010]
- Reranking
  - features of Charniak and Johnson [2005] and Collins [2000]
  - additional language independent features

Grammars with Latent Annotations [Petrov et al., 2006]

Baseline: the Berkeley parser with two versions

- *mainPOS*: only POS tags in preterminals, no morphological annotations
- *fullMorph*: full morphological descriptions

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
mainPOS	72.32	**79.35**	**82.26**	**88.71**	83.84	71.85	**86.75**	**75.19**
fullMorph	**77.82**	79.17	80.22	88.40	**87.18**	**82.28**	85.06	72.82

Lexical Sparsity: Replacing

- Replaced rare words (frequency < 20) by the morphological tag assigned by MarMoT
- Applied the replacement on the Berkeley grammar with main POS tags

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
mainPOS	72.32	79.35	82.26	88.71	83.84	71.85	86.75	75.19
Δ	11.95	0.91	0.73	1.02	5.75	11.22	3.54	1.89
Replace	84.27	80.26	82.99	89.73	89.59	83.07	90.29	77.08

Lexical Sparsity: Extended Lexicon Model [Goldberg and Elhadad, 2013]

- Exploited the available unlabeled data
- Derived tagging probabilities by counting relative frequencies of MarMoT predictions on development and unlabeled data

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
fullMorph	77.82	79.17	80.22	88.40	87.18	82.28	85.06	72.82
Δ	-0.31	0.5	1.32	0.93	1.81	-	3.15	1.75
ExtendLex	77.51	79.67	81.54	89.33	88.99	-	88.21	74.57

Product Grammars [Petrov, 2010]

- Trained 8 PCFG-LAs with different seed values
- Extracted k-best lists and scored them with the product model (Tree-Level inference)

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
ExtendLex	77.51	79.67	81.54	89.33	88.99	-	88.21	74.57
Δ	3.2	1.71	0.59	0.59	1.44	-	3.31	3.64
$Product_{ELex}$	80.71	**81.38**	82.13	**89.92**	90.43	-	91.52	78.21
Replace	84.27	80.26	82.99	89.73	89.59	83.07	90.29	77.08
Δ	1.04	1.03	1.56	0.14	1.13	0.79	1.99	1.58
$Product_{Rep}$	**85.31**	81.29	**84.55**	89.87	**90.72**	**83.86**	**92.28**	**78.66**

Reranking

- Maximum entropy reranker of Charniak and Johnson [2005]
- Language independent features of Charniak and Johnson [2005] and Collins [2000] (*dflt*)
- Additional language independent features this year
  - *dep*: automatic dependency parses of the sentence in question [Farkas and Bohnet, 2012]
  - *Brown* clusters [Brown et al., 1992];
  - *morph*: atomic morphological feature values [Szántó and Farkas, 2014]
    - e.g, features like (Dog-N-Cas=n)

Reranking

- Additional features helped reranking
  - Except a small drop in Swedish *ExtendLex*

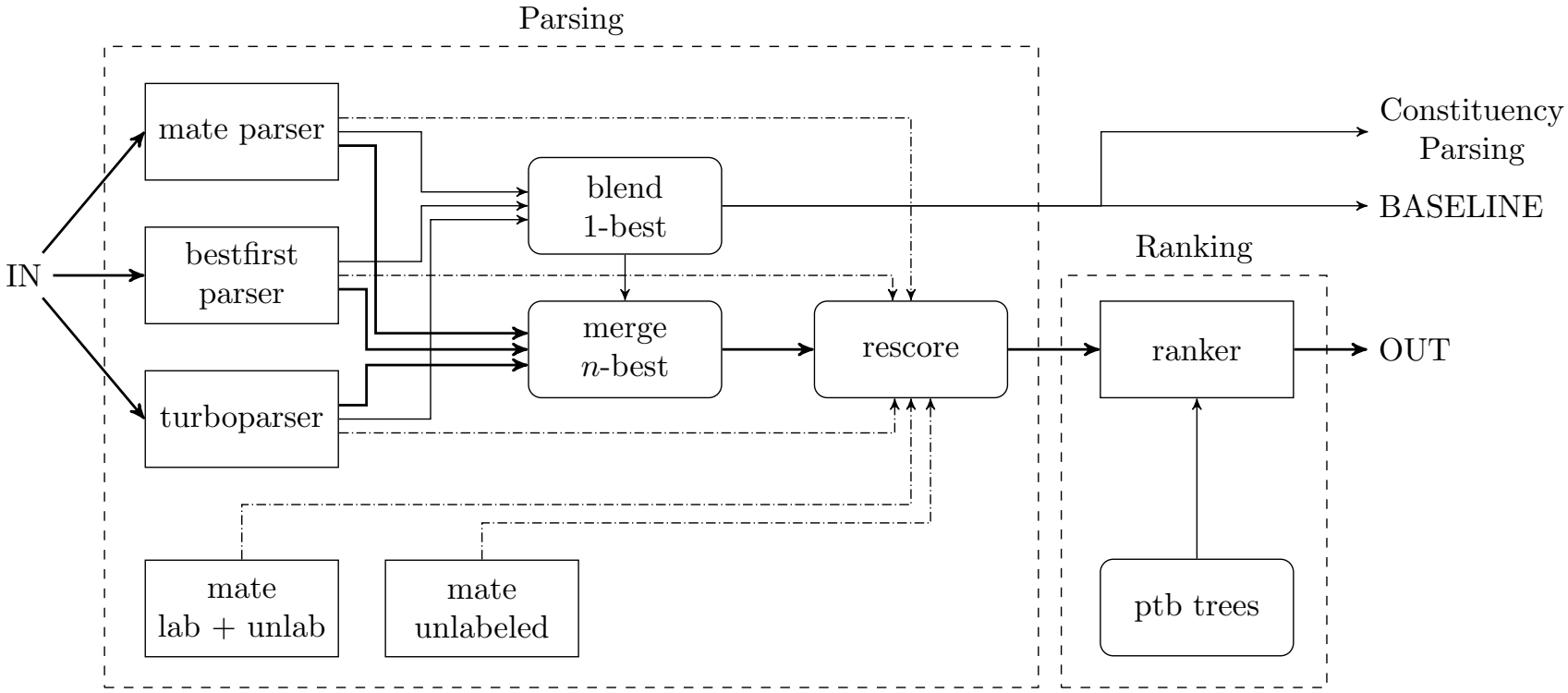
	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
Reranked ExtendLex								
dflt	81.59	81.92	82.83	90.16	91.06	-	89.79	79.09
Δ	2.24	0.84	1.86	0.6	0.99	-	1.65	-0.31
all	83.83	82.76	84.69	**90.76**	**92.05**	-	**91.44**	78.78
Reranked Replace								
dflt	86.11	82.30	84.59	90.02	91.09	83.50	88.31	78.87
Δ	0.62	0.48	1.46	0.45	0.8	1.28	2.22	0.51
dflt + *dep*	**86.73**	**82.78**	**86.05**	90.47	91.89	**84.78**	90.53	**79.38**

all: *dflt* + *morph* + *Brown* + *dep*

Section 3

Dependency Parsing

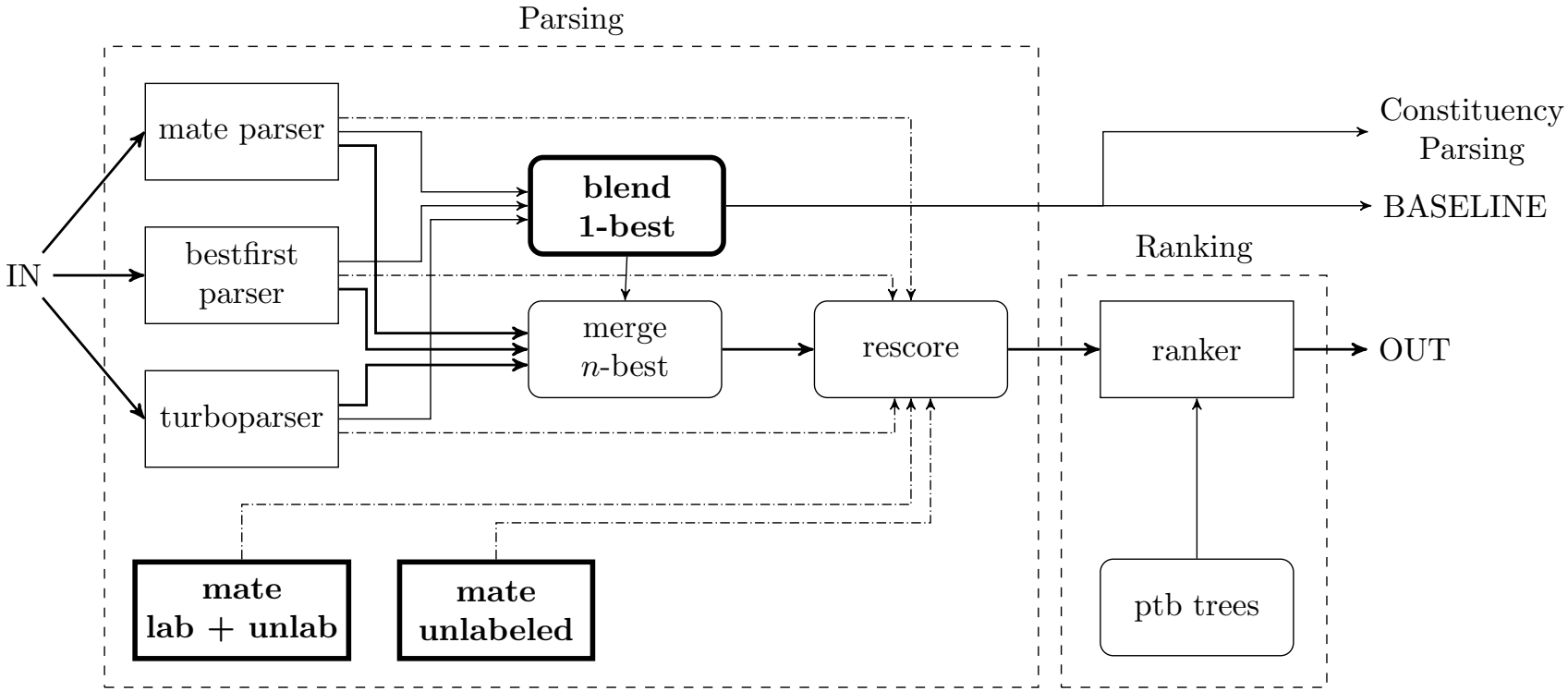
Architecture of the Parsing and Ranking System



- Three parsers:

  mate [Bohnet, 2010], in-house EasyFirst [Goldberg and Elhadad, 2010], TurboParser [Martins et al., 2010]
- Merged list was scored by all three parsers and self-training models [Zhang et al., 2009]
- Scored list was ranked to find the optimal parse

Architecture of the Parsing and Ranking System



- Three parsers:

  mate [Bohnet, 2010], in-house EasyFirst [Goldberg and Elhadad, 2010], TurboParser [Martins et al., 2010]
- Merged list was scored by all three parsers and self-training models [Zhang et al., 2009]
- Scored list was ranked to find the optimal parse

Base parsers

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
mate	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	**85.08**	84.47	**91.69**	**80.05**	83.39	**86.92**	**87.03**	77.18

Supertags [Bangalore and Joshi, 1999]

- Tags that encode syntactic information as features to parsers [Ouchi et al., 2014, Ambati et al., 2014]
- Designed three different models. All of them used:
  - relative position of the head of a word
  - dependency relation of the head
  - relative position of the word's dependents

Model	Additional information	Example (German)
*M*1	-	OC/R+L
*M*2	relations of obligatory dependents of verbs	OC/R+L_OP/L
*M*3	all dependency relations	OC/R+MO/L_OP/L

Supertags

- Applied all supertag models to mate and TurboParser
- All models helped
- Selected the best model for TurboParser and $M1$ for mate

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
turbo	83.98	84.03	91.32	78.99	82.50	86.08	85.27	75.62
Δ	1.1	0.44	0.37	1.06	0.9	0.84	1.76	1.56
turbo$_{stags}$	85.08	84.47	91.69	80.05	83.40	86.92	87.03	77.18
mate	83.96	84.34	91.25	79.66	84.15	85.49	85.96	76.50
Δ	0.78	0.45	0.24	0	0.32	1.03	0.27	0.75
mate$_{stags}$	84.74	84.79	91.49	79.66	84.47	86.52	86.23	77.25

Blending [Sagae and Lavie, 2006]

- Combined parse trees from three base parsers into one
- Added the blended tree into the n-best list if not already there
- The blended tree was also marked to be used as a ranker feature later

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
mate	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	**85.08**	84.47	91.69	80.05	83.39	**86.92**	**87.03**	77.18
Δ_{best}	-0.37	0.32	0.5	0.6	-0.23	-0.09	-0.06	0.98
blend	84.71	**85.10**	**92.19**	**80.65**	84.24	86.83	86.97	**78.23**

Parsing with Unlabeled Data

- Filtered the unlabeled data
  - $5 \leq \textit{length} \leq 20$
  - at most 2 unknown word forms wrt. the training data
  - contain at least one verb (determined by POS)
  - no word forms longer than 20 characters
  - no word forms that have more than 3 punctuation
  - no word forms that occur less than 5 times in the unlabeled data
- Parsed with mate and TurboParser
- Intersected two parser outputs [Sagae and Tsujii, 2007]
- Removed parses when function labels that should occur once occur more often

Parsing with Unlabeled Data

- Two self-trained mate models [McClosky et al., 2006]
  - mate_{ulbl}: 100K sentences of unlabeled data
  - $\text{mate}_{lbl+ulbl}$: treebank training data + unlabeled data as the size of treebank data
- $\text{mate}_{lbl+ulbl}$ is better than mate but worse than TurboParser for Basque and Polish
- But better than all for Swedish

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
mate	84.74	**84.78**	91.49	79.66	**84.47**	86.52	86.23	77.25
turbo	**85.08**	84.47	**91.69**	**80.05**	83.39	**86.92**	**87.03**	77.18
Δ_{mate}	0.28	-0.18	-0.15	0.29	-0.09	-0.19	0.4	0.85
$\text{mate}_{lbl+ulbl}$	85.02	84.60	91.34	79.95	84.38	86.33	86.63	**78.10**
Δ_{mate}	-0.92	-2.35	-3.14	-1.54	-2.21	-0.79	-0.31	-1.77
mate_{ulbl}	83.82	82.43	88.35	78.12	82.26	85.73	85.92	75.48

Ranking

- Same ranking model as for constituencies
- Feature sets for each language were optimized manually via cross-validation on training data

Ranking Features

Last year's features

- Scores from base parsers
- Features indicating if a parse is best according to each model
- Ill-nestedness of trees
- Paths in the constituency tree between head and dependent
- Function label uniqueness

New features

- Scores from mate trained on labeled and unlabeled data
- Features indicating if a tree is the output of the parser blender

Ranking Results

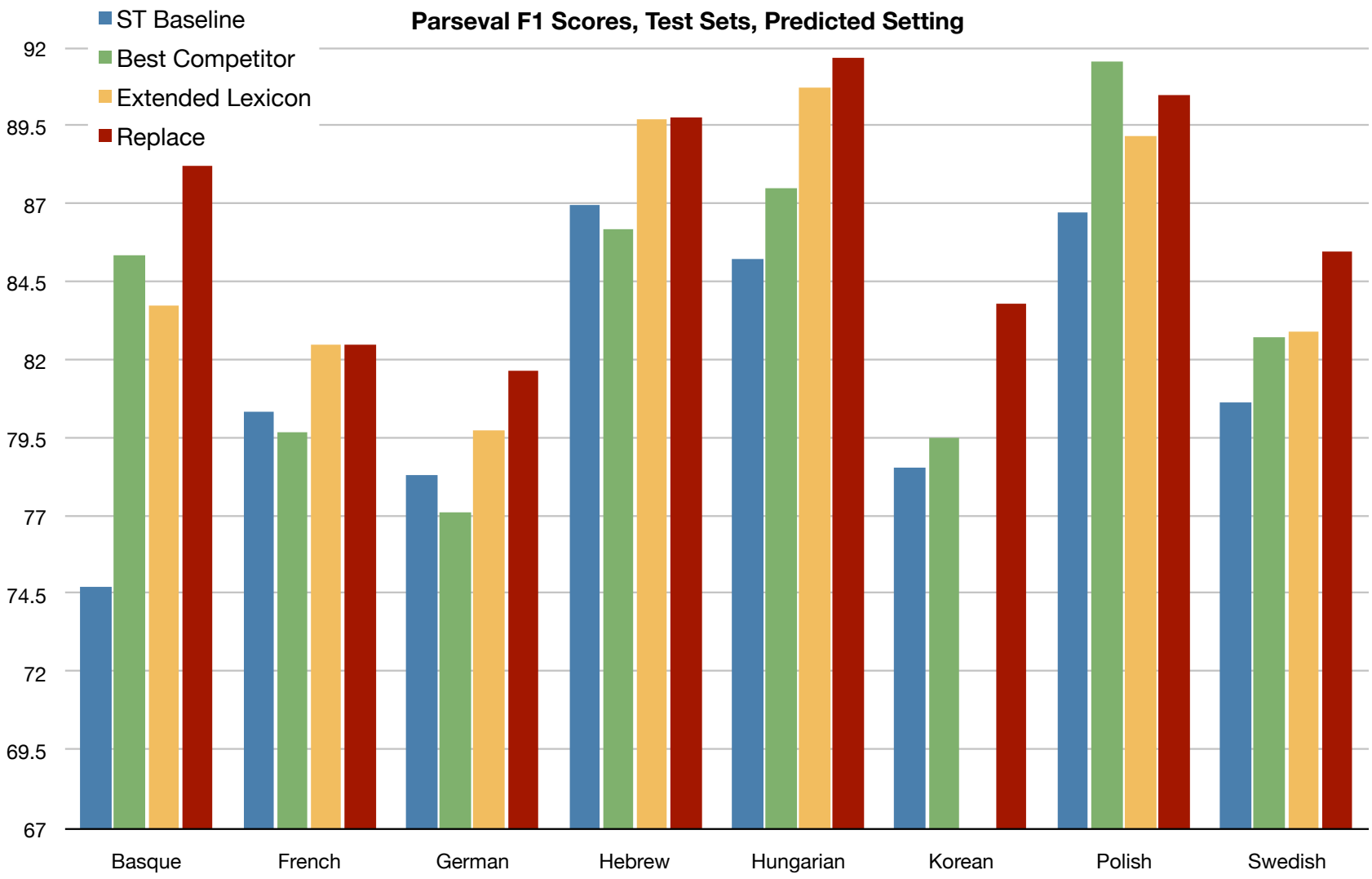
- Ranking improved accuracies on all languages
- Scores from mate trained on labeled and unlabeled data helped across languages
  - with negligible improvements (except Swedish)

	Basque	French	German	Hebrew	Hung.	Korean	Polish	Swedish
Baseline	84.71	85.10	92.19	80.65	84.24	86.83	86.97	78.23
Ranked-dflt	85.80	85.00	92.19	80.49	84.34	87.41	87.59	77.98
Ranked-no-ulbl	86.40	86.00	92.72	**81.96**	84.99	**87.89**	88.00	79.02
$\Delta_{baseline}$	1.75	0.91	0.56	1.28	0.84	1.02	1.1	1.41
Ranked-opt	**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

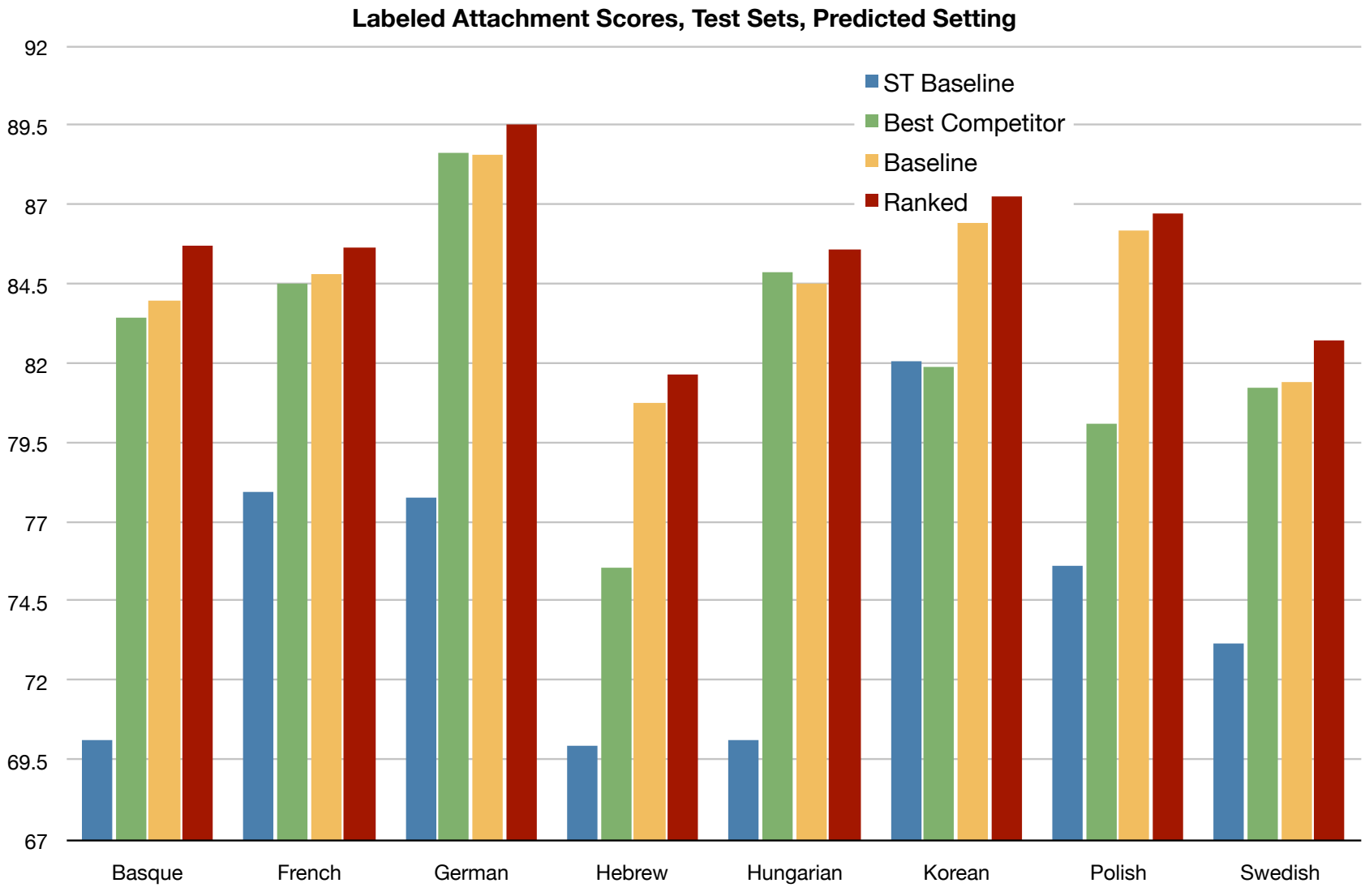
Section 4

Conclusion

Test Set Results – Constituency Parsing



Test Set Results – Dependency Parsing



Conclusions – What We Learned

- Did we improve our last year performance?
  - Lower preprocessing than last year
  - Improved:
    - constituency scores: 4 out of 8 languages
    - dependency scores: 5 out of 8 languages
- Does unlabeled data help?
  - Helps in several ways (e.g. dictionaries, *ExtendLex*) but cannot beat the methods we used last year (e.g. stacking, *Replace*)
  - When utilized as features to (re)rankers, moderately helps
  - An exception is Swedish dependency parsing

Thanks!

Questions?

Ambati, B. R., Deoskar, T., and Steedman, M. (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. Association for Computational Linguistics.

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

Björkelund, A., Çetinoğlu, O., Farkas, R., Müller, T., and Seeker, W. (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. Association for Computational Linguistics.

Bohnet, B. (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. Coling 2010 Organizing Committee.

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):467–479.

Charniak, E. and Johnson, M. (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.

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

Farkas, R. and Bohnet, B. (2012). Stacking of dependency and phrase structure parsers. In *Proceedings of COLING 2012*, pages 849–866, Mumbai, India. The COLING 2012 Organizing Committee.

Goldberg, Y. and Elhadad, M. (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. Association for Computational Linguistics.

Goldberg, Y. and Elhadad, M. (2013). Word Segmentation, Unknown-word Resolution, and Morphological Agreement in a Hebrew Parsing System. *Computational Linguistics*, 39(1):121–160.

Martins, A., Smith, N., Xing, E., Aguiar, P., and Figueiredo, M. (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. Association for Computational Linguistics.

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA. Association for Computational Linguistics.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *In Proceedings of EMNLP*.

Ouchi, H., Duh, K., and Matsumoto, Y. (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. Association for Computational Linguistics.

Petrov, S. (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. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (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.

Sagae, K. and Lavie, A. (2006). Parser combination by reparsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 129–132, New York City, USA. Association for Computational Linguistics.

Sagae, K. and Tsujii, J. (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. Association for Computational Linguistics.

Szántó, Z. and Farkas, R. (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. Association for Computational Linguistics.

Zhang, H., Zhang, M., Tan, C. L., and Li, H. (2009). K-Best Combination of Syntactic Parsers. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1552–1560, Singapore. Association for Computational Linguistics.