UPPSALA UNIVERSITET

How to Parse Low-Resource Languages: Cross-Lingual Parsing, Target Language Annotation, or Both?

Ailsa Meechan-Maddon
Joakim Nivre

Uppsala University, Sweden

Outline

- Aim: to produce a means of parsing low-resource languages
- We compare the usefulness of the following three approaches:
  - **Monolingual:** Trained on small amounts of target language data
  - **Cross-lingual:** Trained only on data from related support languages
  - **Multilingual:** Trained on both support and target language data

Motivation

- Basic NLP technologies are still available only for a tiny fraction of the languages of the world.
- An increasing interest in techniques for supporting low-resource languages.
- Is it more worthwhile to simply annotate a small amount of training data in the target language?

Related Work

- Main approaches:
  - Annotation projection (Hwa et al., 2002; Hwa et al., 2005; Tiedemann, 2014)
  - Model transfer (Zeman and Resnik, 2008; McDonald et al., 2011)
  - Treebank translation (Tiedemann et al., 2014)
  - Multilingual parsing (Ammar et al. 2016, Smith et al. 2018a)

Languages

- From Universal Dependencies v2.3 (Nivre et al., 2016; 2018) we take three language clusters:
  - Faroese
    - **Support languages:** Danish, Norwegian (Nynorsk) and Swedish
  - North Saami
    - **Support languages:** Estonian, Finnish, Hungarian
  - Upper Sorbian
    - **Support languages:** Czech, Polish and Slovak

UPPSALA UNIVERSITET

Methodology

- We adopt a multilingual parsing approach (Ammar et al. 2016, Smith et al. 2018a).
- We use lexicalized models and do not presuppose PoS tagging or any other preprocessing except tokenization for the target language.
- We instead rely on word, character and language embeddings.

UPPSALA UNIVERSITET

Methodology

- We use UUParser v2.3 (de Lhoneux et al., 2017a; Smith et al., 2018a).
- An adaptation of the transition-based parser of Kiperwasser and Goldberg (2016) specifically for multilingual models.
- It relies on a BiLSTM to learn representations of tokens in context.

Experimental setup

- For each language cluster, we train a total of 15 models:
  - **cross-lingual** models on data from every combination of one, two or three support languages (7 models)
  - **multilingual** models on the same data sets plus target language data (7 models)
  - a **monolingual** model only on target language data

Results

	Scandinavian			West Slavic			Uralic	
	−T LAS	+T LAS		−T LAS	+T LAS		−T LAS	+T LAS
Faroese		71.1	Upper Sorbian		58.4	N Saami		58.6
Dan	30.6	74.2	Cze	23.5	64.2	Est	8.5	**60.1**
Nor	35.7	76.2	Pol	34.3	64.2	Fin	7.5	59.5
Swe	24.9	73.9	Slo	29.5	61.9	Hun	4.9	57.5
Dan+Nor	34.5	76.7	Cze+Pol	41.4	63.9	Est+Fin	9.4	55.3
Dan+Swe	35.9	75.5	Cze+Slo	32.6	**65.3**	Est+Hun	8.9	58.4
Nor+Swe	39.6	**77.0**	Pol+Slo	38.8	64.9	Fin+Hun	8.0	57.7
All	44.4	75.3	All	43.3	62.8	All	11.6	56.4

Test set accuracy for target languages (LAS). –T = cross-lingual models trained without target language data. +T = models trained on target language data; monolingual (first row) and multilingual.

Monolingual > Cross-lingual

	Scandinavian −T LAS	Scandinavian +T LAS		West Slavic −T LAS	West Slavic +T LAS		Uralic −T LAS	Uralic +T LAS
Faroese		71.1	Upper Sorbian		58.4	N Saami		58.6
Dan	30.6	74.2	Cze	23.5	64.2	Est	8.5	**60.1**
Nor	35.7	76.2	Pol	34.3	64.2	Fin	7.5	59.5
Swe	24.9	73.9	Slo	29.5	61.9	Hun	4.9	57.5
Dan+Nor	34.5	76.7	Cze+Pol	41.4	63.9	Est+Fin	9.4	55.3
Dan+Swe	35.9	75.5	Cze+Slo	32.6	**65.3**	Est+Hun	8.9	58.4
Nor+Swe	39.6	**77.0**	Pol+Slo	38.8	64.9	Fin+Hun	8.0	57.7
All	44.4	75.3	All	43.3	62.8	All	11.6	56.4

Test set accuracy for target languages (LAS). –T = cross-lingual models trained without target language data. +T = models trained on target language data; monolingual (first row) and multilingual.

UPPSALA UNIVERSITET

Multilingual > Monolingual (most of the time)

	Scandinavian			West Slavic			Uralic	
	–T	+T		–T	+T		–T	+T
	LAS	LAS		LAS	LAS		LAS	LAS
Faroese		71.1	Upper Sorbian		58.4	N Saami		58.6
Dan	30.6	74.2	Cze	23.5	64.2	Est	8.5	**60.1**
Nor	35.7	76.2	Pol	34.3	64.2	Fin	7.5	59.5
Swe	24.9	73.9	Slo	29.5	61.9	Hun	4.9	57.5
Dan+Nor	34.5	76.7	Cze+Pol	41.4	63.9	Est+Fin	9.4	55.3
Dan+Swe	35.9	75.5	Cze+Slo	32.6	**65.3**	Est+Hun	8.9	58.4
Nor+Swe	39.6	**77.0**	Pol+Slo	38.8	64.9	Fin+Hun	8.0	57.7
All	44.4	75.3	All	43.3	62.8	All	11.6	56.4

Test set accuracy for target languages (LAS). –T = cross-lingual models trained without target language data. +T = models trained on target language data; monolingual (first row) and multilingual.

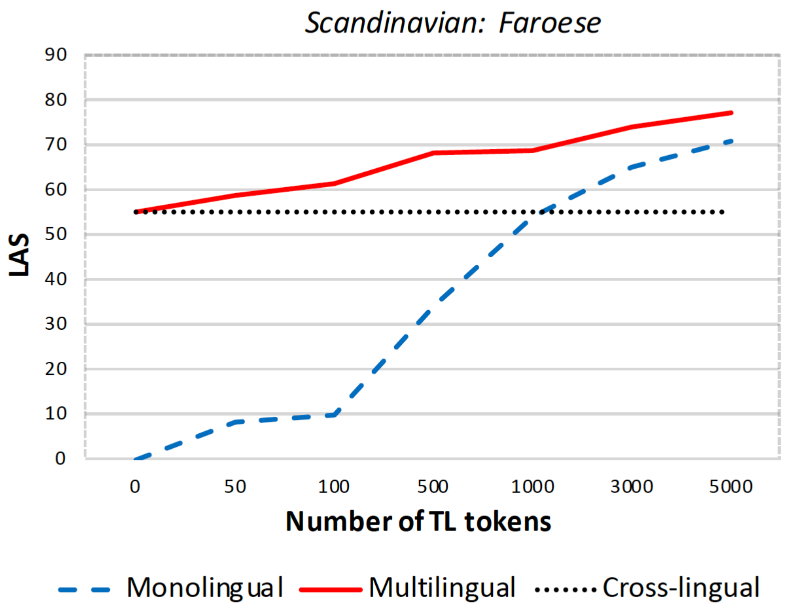
UPPSALA UNIVERSITET

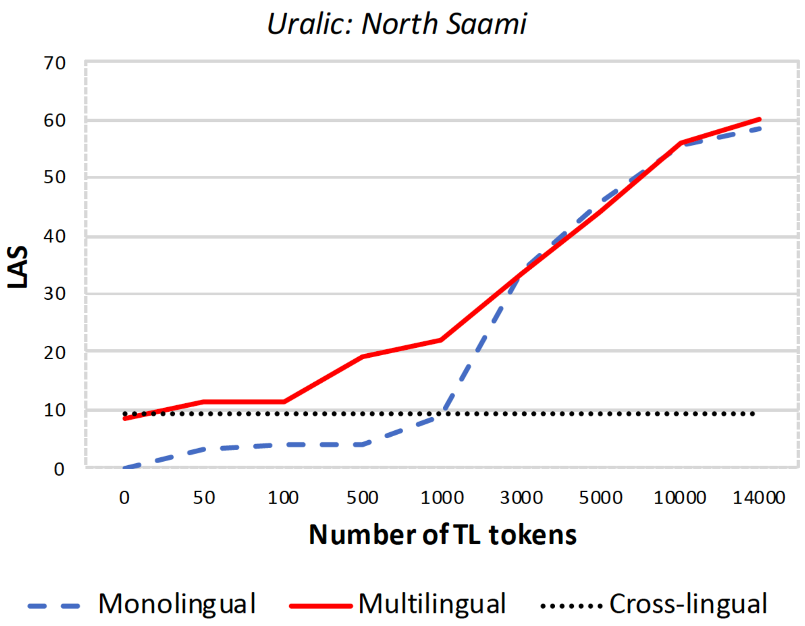
Multilingual > Monolingual (most of the time)

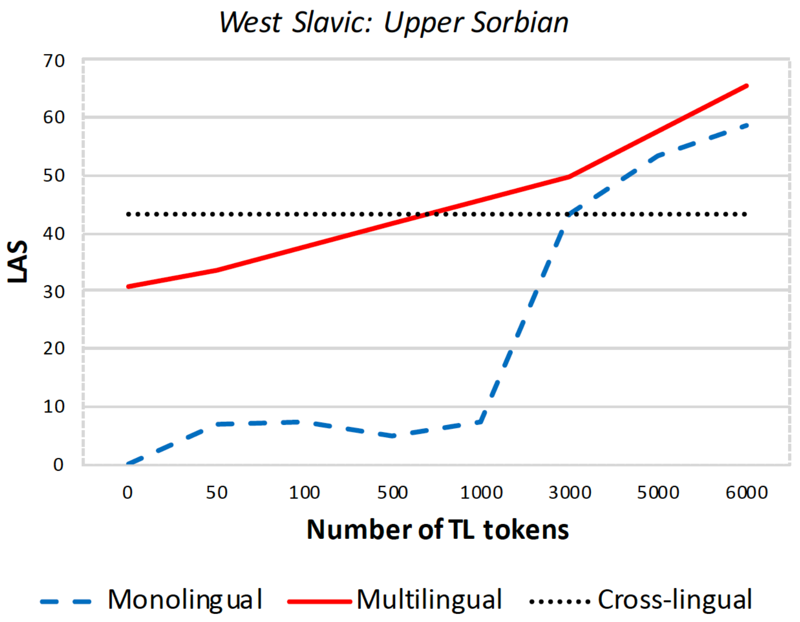
Scandinavian	–T LAS	+T LAS	West Slavic	–T LAS	+T LAS	Uralic	–T LAS	+T LAS
Faroese		71.1	Upper Sorbian		58.4	N Saami		58.6
Dan	30.6	74.2	Cze	23.5	64.2	Est	8.5	**60.1**
Nor	35.7	76.2	Pol	34.3	64.2	Fin	7.5	59.5
Swe	24.9	73.9	Slo	29.5	61.9	Hun	4.9	57.5
Dan+Nor	34.5	76.7	Cze+Pol	41.4	63.9	Est+Fin	9.4	55.3
Dan+Swe	35.9	75.5	Cze+Slo	32.6	**65.3**	Est+Hun	8.9	58.4
Nor+Swe	39.6	**77.0**	Pol+Slo	38.8	64.9	Fin+Hun	8.0	57.7
All	44.4	75.3	All	43.3	62.8	All	11.6	56.4

Test set accuracy for target languages (LAS). –T = cross-lingual models trained without target language data. +T = models trained on target language data; monolingual (first row) and multilingual.

Results: Learning curves







Conclusion

- Training a monolingual model on target language data gives better performance than any cross-lingual model as soon as we have around 200 annotated target language sentences.
- Adding data from related languages to train a multilingual model can improve performance further by up to 7 LAS points.
- In conclusion, to develop a parser for a low-resource language, annotate as much data as you can afford and add data from related languages if available.

UPPSALA UNIVERSITET

Bibliography

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. Transactions of the Association for Computational Linguistics, 4:431–444.
- Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre.
- 2017a. From raw text to Universal Dependencies – Look, no tags! In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 207–217.
- Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 20–30.
- Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A. Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. Natural Language Engineering, 24:91–122.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 392–399.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. Natural Language Engineering, 11(3):311–325.
- Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. Transactions of the Association for Computational Linguistics, 4:313–327.
- Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), pages 1403–1414.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 62–72.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajiˇc, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC).
- Joakim Nivre et al., 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018a. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In Proceedings of the 2018 CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.
- Jörg Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In Proceedings of the 25th International Conference on Computational Linguistics (COLING), pages 1854–1864.
- Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 35–42.

Resources

Language	Treebank	Train	Dev	Test
Faroese	OFT	4.9k	2.5k	2.5k
Danish	DDT	80k		
Norwegian	Nynorsk	245k		
Swedish	Talbanken	67k		
Upper Sorbian	UFAL	5.8k	2.7k	2.7k
Czech	PDT	300k		
Polish	LFG	105k		
	SZ	63k		
Slovak	SNK	81k		
North Saami	Giella	14.3k	2.5k	10k
Estonian	EDT	288k		
Finnish	FTB	128k		
	TDT	163k		
Hungarian	Szeged	20k		

Table 1: Data sets (UD v2.3) with number of tokens.

Resources

	Scandinavian					West Slavic					Uralic			
	−Target		+Target			−Target		+Target			−Target		+Target	
	UAS	LAS	UAS	LAS		UAS	LAS	UAS	LAS		UAS	LAS	UAS	LAS
Faroese			78.6	71.1	Upper Sorbian			66.0	58.4	N Saami			66.0	58.6
Dan	45.9	30.6	81.5	74.2	Cze	33.1	23.5	71.5	64.2	Est	22.4	8.5	**68.8**	**60.1**
Nor	47.4	35.7	84.0	76.2	Pol	44.9	34.3	71.3	64.2	Fin	22.5	7.5	67.3	59.5
Swe	45.9	24.9	81.1	73.9	Slo	41.2	29.5	68.6	61.9	Hun	19.4	4.9	65.6	57.5
Dan+Nor	48.5	34.5	83.7	76.7	Cze+Pol	51.1	41.4	72.2	63.9	Est+Fin	24.7	9.4	64.5	55.3
Dan+Swe	55.9	35.9	82.7	75.5	Cze+Slo	44.1	32.6	**72.5**	**65.3**	Est+Hun	23.8	8.9	67.0	58.4
Nor+Swe	56.4	39.6	**83.9**	**77.0**	Pol+Slo	47.5	38.8	72.2	64.9	Fin+Hun	20.8	8.0	65.9	57.7
All	57.7	44.4	82.8	75.3	All	52.4	43.3	69.6	62.8	All	27.1	11.6	65.4	56.4

Table 2: Test set accuracy for target languages (UAS, LAS). −Target = cross-lingual models trained without target language data. +Target = models trained on target language data; monolingual (first row) and multilingual.