Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning — Supplementary material —

1 Resources and Features

1.1 Corpus

Table 1 describes the different languages considered in our experiments and the different corpora used. Evaluation is carried out on the test sets of the treebanks of the Universal Dependency Treebank Project (UDT v2.0 in Table 1) [3], for Czech on the CoNLL'09 Shared Task on Dependency Parsing [1], for Greek on the Greek Dependency Treebank [6] used in the the CoNLL'07 Shared Task on Dependency Parsing [4] and for Arabic on the Arabic Treebank [2]. In the Universal Dependency Treebank, all data are annotated with the Universal POS tagset of [5]. For the other three corpora, as well as for the English side of all parallel corpora, the tagset was mapped to the universal POS tagset using the mappings from [5]. Note that for Greek, only the the test set of the treebank is freely available.

	Language	Language Family	Parallel Corpus	Labeled Data
ar	Arabic	Afro-Asiatic/Semitic	NIST	Arabic Treebank
cs	Czech	Indo-European/Balto-Slavic	Europarl	CoNLL 2009 Shared Task
de	German	Indo-European/Germanic	Europarl	UDT v2.0
el	Greek	Indo-European/Hellenic	Europarl	Greek Dependency Treebank
es	Spanish	Indo-European/Italic	Europarl	UDT v2.0
fi	Finnish	Uralic/Finnic	Europarl	UDT v2.0
fr	French	Indo-European/Italic	Europarl	UDT v2.0
id	Indonesian	Austronesian/Malayo-Polynesian	Open Subtitle	UDT v2.0
it	Italian	Indo-European/Italic	Europarl	UDT v2.0
sv	Swedish	Indo-European/Germanic	Europarl	UDT v2.0

Table 1: Description of the different languages considered and the resources considered for each language. UDT stands for Universal Dependency Treebank.

1.2 Features

We use, in all our models, a standard feature set, similar to the one used in previous works that is made of the following features:

- for the current word, as well as for the two previous and the two following words:
  - lowercased word form if the word appears more than 10 times in the training set;
  - last 2 and 3 letters of the word if they appear in more than 20 different word types;
- two binary features that indicates whether the word starts with an uppercase or not;
- two binary features that that indicates whether the the word contains an hyphen;
- two binary features that indicates whether the word is written in Greek or Latin alphabet;
- the previous two labels predicted, the conjunction of both and the conjunction of the previous label and the previous word are

As sole preprocessing, all digits are mapped to a special token.

2 Partially observed CRF model

We use our own implementation of the partially observed CRF model $\hat{\mathcal{Y}}^{\text{CRF}}_{\text{wik.}}$+L described in [7], with the feature templates depicted above.[1] For each language, we sample 100 000 sentences and run 30 iterations of resilient backpropagation (R-Prop) algorithm with elastic net regularization. In the context of under resourced languages, one would be unable to use a development set to fine tune the hyper-parameters, so following [7] we arbitrarily set the ℓ_1 and ℓ_2 regularization parameters to 1.

References

[1] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL'09, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[1] In this case, we do not filter the word forms or the suffixes, but use ℓ_1 regularization to perform feature selection. We found this to be slightly better than using only ℓ_2 regularization as in [7].

[2] Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. The Penn Arabic Treebank: Building a large-scale annotated arabic corpus. In *Proceedings of the Conference on Arabic Language Resources and Tools*, 2004.

[3] Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Short Papers*, ACL'13, pages 92–97, Sofia, Bulgaria, August 2013.

[4] Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[5] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC'12, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[6] Prokopis Prokopidis, Elina Desypri, Maria Koutsombogera, Haris Papageorgiou, and Stelios Piperidis. Theoretical and practical issues in the construction of a greek dependency treebank. In Montserrat Civit, Sandra Kubler, and Ma. Antonia Marti, editors, *Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories*, TLT'05, pages 149–160, Barcelona, Spain, December 2005. Universitat de Barcelona.

[7] Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12, 2013.