Distant Supervision from Disparate Sources
for Low-Resource Part-of-Speech Tagging

Barbara Plank and **Željko Agić**
Department of Computer Science
IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
`bplank@itu.dk, zeag@itu.dk`

Abstract

We introduce DsDs: a cross-lingual neural part-of-speech tagger that learns from disparate sources of distant supervision, and realistically scales to hundreds of low-resource languages. The model exploits annotation projection, instance selection, tag dictionaries, morphological lexicons, and distributed representations, all in a uniform framework. The approach is simple, yet surprisingly effective, resulting in a new state of the art *without* access to any gold annotated data.

1 Introduction

Low-resource languages lack manually annotated data to learn even the most basic models such as part-of-speech (POS) taggers. To compensate for the absence of direct supervision, work in cross-lingual learning and distant supervision has discovered creative use for a number of alternative data sources to learn feasible models:

– aligned *parallel corpora* to project POS annotations to target languages (Yarowsky et al., 2001; Agić et al., 2015; Fang and Cohn, 2016),
– noisy *tag dictionaries* for type-level approximation of full supervision (Li et al., 2012),
– combination of projection and *type constraints* (Das and Petrov, 2011; Täckström et al., 2013),
– rapid annotation of *seed training data* (Garrette and Baldridge, 2013; Garrette et al., 2013).

However, only one or two compatible sources of distant supervision are typically employed. In reality severely under-resourced languages may require a more pragmatic "take what you can get" viewpoint. Our results suggest that combining supervision sources is the way to go about creating viable low-resource taggers.

We propose a method to strike a balance between model simplicity and the capacity to easily integrate heterogeneous learning signals. Our
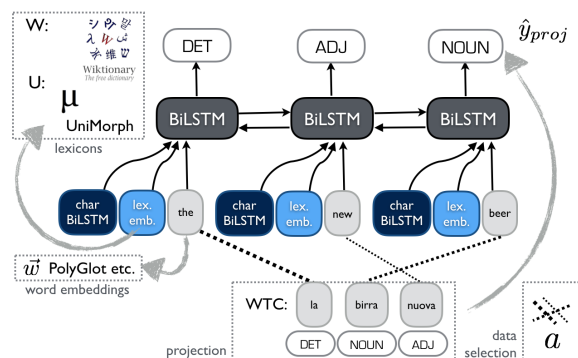

Figure 1: Illustration of DsDs (Distant Supervision from Disparate Sources).

system is a uniform neural model for POS tagging that learns from *disparate sources of distant supervision* (DsDs). We use it to combine: i) multi-source annotation projection, ii) instance selection, iii) noisy tag dictionaries, and iv) distributed word and sub-word representations. We examine how far we can get by exploiting only the wide-coverage resources that are currently readily available for more than 300 languages, which is the breadth of the parallel corpus we employ.

DsDs yields a new state of the art by jointly leveraging disparate sources of distant supervision in an experiment with 25 languages. We demonstrate: i) substantial gains in carefully selecting high-quality instances in annotation projection, ii) the usefulness of lexicon features for neural tagging, and iii) the importance of word embeddings initialization for faster convergence.

2 Method

DsDs is illustrated in Figure 1. The base model is a bidirectional long short-term memory network (bi-LSTM) (Graves and Schmidhuber, 2005; Hochreiter and Schmidhuber, 1997; Plank et al., 2016; Kiperwasser and Goldberg, 2016). Let $x_{1:n}$

be a given sequence of input vectors. In our base model, the input sequence consists of word embeddings $\vec{w}$ and the two output states of a character-level bi-LSTM $\vec{c}$. Given $x_{1:n}$ and a desired index i, the function $BiRNN_\theta(x_{1:n}, i)$ (here instantiated as LSTM) reads the input sequence in forward and reverse order, respectively, and uses the concatenated ($\circ$) output states as input for tag prediction at position i.[1] Our model differs from prior work on the type of input vectors $x_{1:n}$ and distant data sources, in particular, we extend the input with lexicon embeddings, all described next.

Annotation projection. Ever since the seminal work of Yarowsky et al. (2001), projecting sequential labels from source to target languages has been one of the most prevalent approaches to cross-lingual learning. Its only requirement is that parallel texts are available between the languages, and that the source side is annotated for POS.

We apply the approach by Agić et al. (2016), where labels are projected from multiple sources and then decoded through weighted majority voting with word alignment probabilities and source POS tagger confidences. We exploit their wide-coverage Watchtower corpus (WTC), in contrast to the typically used Europarl data. Europarl covers 21 languages of the EU with 400k-2M sentence pairs, while WTC spans 300+ widely diverse languages with only 10-100k pairs, in effect sacrificing depth for breadth, and introducing a more radical domain shift. However, as our results show little projected data turns out to be the most beneficial, reinforcing breadth for depth.

While Agić et al. (2016) selected 20k projected sentences at random to train taggers, we propose a novel alternative: selection by *coverage*. We rank the target sentences by percentage of words covered by word alignment from 21 sources of Agić et al. (2016), and select the top k covered instances for training. In specific, we employ the mean coverage ranking of target sentences, whereby each target sentence is coupled with the arithmetic mean of the 21 individual word alignment coverages for each of the 21 source-language sentences. We show that this simple approach to instance selection offers substantial improvements: across all languages, we learn better taggers with significantly fewer training instances.

Dictionaries. Dictionaries are a useful source for distant supervision (Li et al., 2012; Täckström et al., 2013). There are several ways to exploit such information: i) as type constraints during encoding (Täckström et al., 2013), ii) to guide unsupervised learning (Li et al., 2012), or iii) as additional signal at training. We focus on the latter and evaluate two ways to integrate lexical knowledge into neural models, while comparing to the former two: a) by representing lexicon properties as n-hot vector (e.g., if a word has two properties according to lexicon src, it results in a 2-hot vector, if the word is not present in src, a zero vector), with m the number of lexicon properties; b) by *embedding* the lexical features, i.e., $\vec{e}_{src}$ is a lexicon src embedded into an l-dimensional space. We represent $\vec{e}_{src}$ as concatenation of all embedded m properties of length l, and a zero vector otherwise. Tuning on the dev set, we found the second embedding approach to perform best, and simple concatenation outperformed mean vector representations.

We evaluate two dictionary sources, motivated by ease of accessibility to many languages: WIKTIONARY, a word type dictionary that maps tokens to one of the 12 Universal POS tags (Li et al., 2012; Petrov et al., 2012); and UNIMORPH, a morphological dictionary that provides inflectional paradigms across 350 languages (Kirov et al., 2016). For Wiktionary, we use the freely available dictionaries from Li et al. (2012) and Agić et al. (2017). The size of the dictionaries ranges from a few thousands (e.g., Hindi and Bulgarian) to 2M (Finnish UniMorph). Sizes are provided in Table 1, first columns. UniMorph covers between 8-38 morphological properties (for English and Finnish, respectively).

Word embeddings. Embeddings are available for many languages. Pre-initialization of $\vec{w}$ offers consistent and considerable performance improvements in our distant supervision setup (Section 4). We use off-the-shelf Polyglot embeddings (Al-Rfou et al., 2013), which performed consistently better than FastText (Bojanowski et al., 2016).

3 Experiments

Baselines. We compare to the following weakly-supervised POS taggers:
– **AGIC:** Multi-source annotation projection with Bible parallel data by Agić et al. (2015).
– **DAS:** The label propagation approach by Das and Petrov (2011) over Europarl data.

---

[1]CRF decoding did not consistently improve POS accuracy, as recently also independently found (Yang et al., 2018).

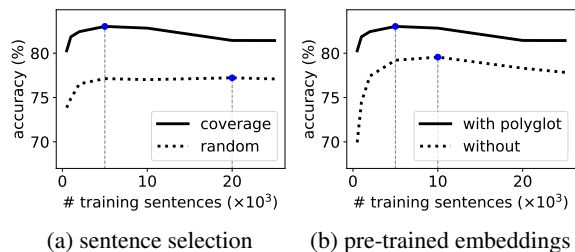(a) sentence selection   (b) pre-trained embeddings

Figure 2: Learning curves for: a) random vs. coverage-based sentence selection in annotation projection, both with Polyglot embeddings, and b) pre-trained embeddings on top of coverage-based selection. Means over 21 languages.

– **GARRETTE:** The approach by Garrette and Baldridge (2013) that works with projections, dictionaries, and unlabeled target text.
– **LI:** Wiktionary supervision (Li et al., 2012).

Data. Our set of 25 languages is motivated by accessibility to embeddings and dictionaries. In all experiments we work with the 12 Universal POS tags (Petrov et al., 2012). For development, we use 21 dev sets of the Universal Dependencies 2.1 (Nivre et al., 2017). We employ UD test sets on additional languages as well as the test sets of Agić et al. (2015) to facilitate comparisons. Their test sets are a mixture of CoNLL (Buchholz and Marsi, 2006; Nivre et al., 2007) and HamleDT test data (Zeman et al., 2014), and are more distant from the training and development data.

Model and parameters. We extend an off-the-shelf state-of-the-art bi-LSTM tagger with lexicon information. The code is available at: `https://github.com/bplank/bilstm-aux`. The parameter $l=40$ was set on dev data across all languages. Besides using 10 epochs, word dropout rate ($p=.25$) and 40-dimensional lexicon embeddings, we use the parameters from Plank et al. (2016). For all experiments, we average over 3 randomly seeded runs, and provide mean accuracy. For the learning curve, we average over 5 random samples with 3 runs each.

4 Results

Table 1 shows the tagging accuracy for individual languages, while the means over all languages are given in Figure 2. There are several take-aways.

Data selection. The first take-away is that coverage-based instance selection yields substan-

tially better training data. Most prior work on annotation projection resorts to arbitrary selection; informed selection clearly helps in this noisy data setup, as shown in Figure 2 (a). Training on 5k instances results in a sweet spot; more data (10k) starts to decrease performance, at a cost of runtime. Training on all WTC data (around 120k) is worse for most languages. From now on we consider the 5k model trained with Polyglot as our baseline (Table 1, column "5k"), obtaining a mean accuracy of 83.0 over 21 languages.

Embeddings initialization. Polyglot initialization offers a large boost; on average +3.8% absolute improvement in accuracy for our 5k training scheme, as shown in Figure 2 (b). The big gap in low-resource setups further shows their effectiveness, with up to 10% absolute increase in accuracy when training on only 500 instances.

Lexical information. The main take-away is that lexical information helps neural tagging, and *embedding* it proves the most helpful. Embedding Wiktionary tags reaches 83.7 accuracy on average, versus 83.4 for n-hot encoding, and 83.2 for type constraints. Only on 4 out of 21 languages are type constraints better. This is the case for only one language for n-hot encoding (French). The best approach is to embed both Wiktionary and Unimorph, boosting performance further to 84.0, and resulting in our final model. It helps the most on morphological rich languages such as Uralic.

On the test sets (Table 4, right) DsDs reaches 87.2 over 8 test languages intersecting Li et al. (2012) and Agić et al. (2016). It reaches 86.2 over the more commonly used 8 languages of Das and Petrov (2011), compared to their 83.4. This shows that our novel "soft" inclusion of noisy dictionaries is superior to a hard decoding restriction, and including lexicons in neural taggers helps. We did not assume any gold data to further enrich the lexicons, nor fix possible tagset divergences.

5 Discussion

Analysis. The inclusion of lexicons results in higher coverage and is part of the explanation for the improvement of DsDs; see correlation in Figure 3 (a). What is more interesting is that our model benefits from the lexicon beyond its content: OOV accuracy for words *not* present in the lexicon overall improves, besides the expected improvement on known OOV, see Figure 3 (b).

Language	LEX (10^3)		DEV SETS (UD2.1)					TEST SETS				
	W	U	5k	TC_W	$n\text{-hot}_W$	$\vec{e}_W$	DsDs		DAS	LI	GARRETTE	AGIC	DsDs
Bulgarian (bg)	3	47	88.6	88.6	88.9	89.6	**89.7**		–	–	83.1	77.7	**83.9**
Croatian (hr)	20	–	84.9	**85.4**	84.9	84.8	†84.8		–	–	–	67.1	†**78.0**
Czech (cs)	14	72	86.6	86.6	86.9	**87.6**	87.2		–	–	–	73.3	**86.8**
Danish (da)	22	24	89.6	89.0	89.8	**90.2**	90.0		83.2	83.3	78.8	79.0	**84.5**
Dutch (nl)	52	26	88.3	88.9	89.0	89.7	**89.8**		79.5	**86.3**	–	–	83.9
English (en)	358	91	86.5	**87.4**	86.8	87.3	87.3		–	**87.1**	80.8	73.0	85.7
Finnish (fi)	104	2,345	81.5	81.2	81.8	82.4	**82.4**		–	–	–	–	–
French (fr)	17	274	91.0	89.6	**91.7**	91.2	91.4		–	–	85.5	76.6	**88.7**
German (de)	62	71	85.0	86.4	85.5	86.0	**86.7**		82.8	85.8	**87.1**	80.2	84.1
Greek (el)	21	–	80.6	**85.7**	80.2	80.5	†80.5		82.5	79.2	64.4	52.3	†**81.1**
Hebrew (he)	3	12	76.0	**76.1**	75.5	74.9	75.3		–	–	–	–	–
Hindi (hi)	2	26	64.6	64.6	64.8	65.4	**66.2**		–	–	–	67.6	63.1
Hungarian (hu)	13	13	75.6	75.6	75.3	75.7	**77.9**		–	–	**77.9**	72.0	77.3
Italian (it)	478	410	91.9	91.7	93.4	93.5	**93.7**		86.8	86.5	83.5	76.9	**92.1**
Norwegian (no)	47	18	90.9	90.9	90.9	91.0	**91.5**		–	–	84.3	76.7	**86.2**
Persian (fa)	4	26	42.8	43.0	43.7	43.5	**43.8**		–	–	–	59.6	43.6
Polish (pl)	6	132	84.7	84.6	84.2	84.8	**86.0**		–	–	–	75.1	**84.4**
Portuguese	41	211	91.4	91.5	92.3	**92.9**	92.2		87.9	84.5	87.3	83.8	**89.4**
Romanian (ro)	7	4	83.9	83.9	84.8	85.3	**86.3**		–	–	–	–	–
Spanish (es)	234	324	90.4	88.6	91.0	91.5	**92.0**		84.2	86.4	88.7	81.4	**91.7**
Swedish (sv)	89	67	88.9	88.9	89.6	**89.9**	89.9		80.5	**86.1**	76.1	75.2	83.1
AVG(21)			83.0	83.2	83.4	83.7	**84.0**	AVG(8: DAS)	83.4	84.8	80.8	75.5	**86.2**
								AVG(8: LI∩AGIC)	–	84.9	80.8	75.2	**87.2**
GERMANIC (6)			88.2	88.6	88.6	89.0	**89.2**	GERMANIC (4: DAS)	81.5	**85.4**	–	–	83.9
ROMANCE (5)			89.7	89.0	90.6	90.9	**91.1**	ROMANCE (3: DAS)	86.3	85.8	86.5	80.7	**91.1**
SLAVIC (4)			86.2	86.3	86.2	86.7	**86.9**						
INDO-IRANIAN (2)			53.7	53.8	54.3	54.4	**55.0**						
URALIC (2)			78.5	78.4	78.6	79.0	**80.1**						

Table 1: Results on the development sets and comparison of our best model to prior work. LEX: Size (word types) of dictionaries (W: Wiktionary, U: UniMorph). TC_W: type-constraints using Wiktionary; $\vec{e}_W$ (embedded Wiktionary tags), DsDs: our model with $\vec{e}_{W\cup U}$. Results indicated by † use W only. Best result in boldface; in case of equal means, the one with lower std is boldfaced. Averages over language families (with two or more languages in the sample, number of languages in parenthesis).

More languages. All data sources employed in our experiment are very high-coverage. However, for true low-resource languages, we cannot safely assume the availability of all disparate information sources. Table 2 presents results for four additional languages where some supervision sources are missing. We observe that adding lexicon information always helps, even in cases where only 1k entries are available, and embedding it is usually the most beneficial way. For closely-related languages such as Serbian and Croatian, using resources for one aids tagging the other, and modern resources are a better fit. For example, using the Croatian WTC projections to train a model for Serbian is preferable over in-language Serbian Bible data where the OOV rate is much higher.

How much gold data? We assume not having access to *any* gold annotated data. It is thus interesting to ask how much gold data is needed to reach our performance. This is a tricky question, as training within the same corpus naturally favors the same corpus data. We test both in-corpus (UD)
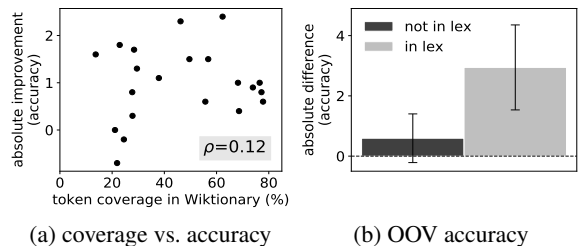

(a) coverage vs. accuracy (b) OOV accuracy

Figure 3: Analysis of DsDs accuracy improvements over the baseline on all development languages with respect to a) token coverage by the lexicon, including Pearson's ρ; b) OOV accuracy for tokens in/not in the lexicon, with 95% confidence intervals of the mean. Here, a token is covered if we can find it in at least one lexicon.

and out-of-corpus data (our test sets) and notice an important gap: while in-corpus only 50 sentences are sufficient, outside the corpus one would need over 200 sentences. This experiment was done for a subset of 18 languages with both in- and out-of-corpus test data.

LANGUAGE	TEST	PROJ	EMB	LEX (10^3)		TEST SETS					
				W	U	TnT	5k	TC_W	n-hot$_W$	$\vec{e}_W$	DsDs
Basque (eu)	UD	Bible	eu	1	–	57.5	61.8	61.8	61.4	**62.7**	†**62.7**
Basque (eu)	CoNLL	Bible	eu	1	–	57.0	60.3	60.3	60.3	**61.3**	†**61.3**
Estonian (et)	UD	WTC	et	–	10	79.5	80.6	–	–	–	**81.5**
Serbian (sr)	UD	WTC (hr)	hr	(hr) 20	–	84.0	84.7	**85.5**	85.1	85.2	†85.2
Serbian (sr)	UD	Bible (sr)	hr	(hr) 20	–	77.1	78.9	79.4	80.5	**80.7**	†**80.7**
Tamil (ta)	UD	WTC	ta	–	–	58.2	**61.2**	–	–	–	–

Table 2: Results for languages with missing data sources: WTC projections, Wiktionary (W), or Uni-Morph (U). Test sets (TEST), projection sources (PROJ), and embeddings languages (EMB) are indicated. Comparison to TnT (Brants, 2000) trained on PROJ. Results indicated by † use W only.

Further comparison. In Table 1 we directly report the accuracies from the original contributions by DAS, LI, GARRETTE, and AGIĆ over the same test data. We additionally attempted to reach the scores of LI by running their tagger over the Table 1 data setup. The results are depicted in Figure 4 as mean accuracies over EM iterations until convergence. We show: i) LI peaks at 10 iterations for their test languages, and at 35 iterations for all the rest. This is in slight contrast to 50 iterations that Li et al. (2012) recommend, although selecting 50 does not dramatically hurt the scores; ii) Our replication falls ∼5 points short of their 84.9 accuracy. There is a large 33-point accuracy gap between the scores of Li et al. (2012), where the dictionaries are large, and the other languages in Figure 4, with smaller dictionaries.

Compared to DAS, our tagger clearly benefits from pre-trained word embeddings, while theirs relies on label propagation through Europarl, a much cleaner corpus that lacks the coverage of the noisier WTC. Similar applies to Täckström et al. (2013), as they use 1-5M near-perfect parallel sentences. Even if we use much smaller and noisier data sources, DsDs is almost on par: 86.2 vs. 87.3 for the 8 languages from Das and Petrov (2011), and we even outperform theirs on four languages: Czech, French, Italian, and Spanish.

6 Related Work

Most successful work on low-resource POS tagging is based on projection (Yarowsky et al., 2001), tag dictionaries (Li et al., 2012), annotation of seed training data (Garrette and Baldridge, 2013) or even more recently some combination of these, e.g., via multi-task learning (Fang and
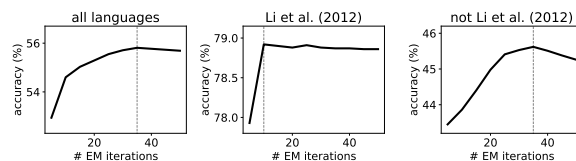

Figure 4: The performance of LI with our dictionary data over EM iterations, separate for the languages from Li et al. (2012) and all the remaining languages in Table 1.

Cohn, 2016; Kann et al., 2018). Our paper contributes to this literature by leveraging a range of prior directions in a unified, neural test bed.

Most prior work on neural sequence prediction follows the commonly perceived wisdom that hand-crafted features are unnecessary for deep learning methods. They rely on end-to-end training without resorting to additional linguistic resources. Our study shows that this is not the case. Only few prior studies investigate such sources, e.g., for MT (Sennrich and Haddow, 2016; Chen et al., 2017; Li et al., 2017; Passban et al., 2018) and Sagot and Martínez Alonso (2017) for POS tagging use lexicons, but only as n-hot features and without examining the cross-lingual aspect.

7 Conclusions

We show that our approach of distant supervision from disparate sources (DsDs) is simple yet surprisingly effective for low-resource POS tagging. Only 5k instances of projected data paired with off-the-shelf embeddings and lexical information integrated into a neural tagger are sufficient to reach a new state of the art, and both data selection and embeddings are essential components to boost neural tagging performance.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Željko Agić, Barbara Plank, and Anders Søgaard. 2017. Cross-lingual tagger evaluation without test data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 248–253. Association for Computational Linguistics.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics.

Huadong Chen, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1936–1945. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 178–186, Berlin, Germany. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of pos-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592, Sofia, Bulgaria. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Katharina Kann, Johannes Bjerva, Isabelle Augenstein, Barbara Plank, and Anders Sgaard. 2018. Character-level supervision for low-resource pos tagging. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 1–11, Melbourne. Association for Computational Linguistics.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *arXiv preprint arXiv:1603.04351*.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou. 2017. Modeling source syntax for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–697, Vancouver, Canada. Association for Computational Linguistics.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods*

in Natural Language Processing and Computational Natural Language Learning, pages 1389–1398, Jeju Island, Korea. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, and Lars Ahrenberg et al. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. 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. Association for Computational Linguistics.

Peyman Passban, Qun Liu, and Andy Way. 2018. Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 58–68. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Benoît Sagot and Héctor Martínez Alonso. 2017. Improving neural tagging with lexical information. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 25–31, Pisa, Italy. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

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

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 1–8. Association for Computational Linguistics.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. Hamledt: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.