Improving neural tagging with lexical information

Benoît Sagot and **Héctor Martínez Alonso**
Inria
Paris, France
{benoit.sagot,hector.martinez-alonso}@inria.fr

Abstract

Neural part-of-speech tagging has achieved competitive results with the incorporation of character-based and pre-trained word embeddings. In this paper, we show that a state-of-the-art bi-LSTM tagger can benefit from using information from morphosyntactic lexicons as additional input. The tagger, trained on several dozen languages, shows a consistent, average improvement when using lexical information, even when also using character-based embeddings, thus showing the complementarity of the different sources of lexical information. The improvements are particularly important for the smaller datasets.

1 Introduction

Part-of-speech tagging is now a classic task in natural language processing. Its aim is to associate each "word" with a morphosyntactic tag, whose granularity can range from a simple morphosyntactic category, or part-of-speech (hereafter PoS), to finer categories enriched with morphological features (gender, number, case, tense, mood, person, etc.).

The use of machine learning algorithms trained on manually annotated corpora has long become the standard way to develop PoS taggers. A large variety of algorithms have been used, such as (in approximative chronological order) bigram and trigram hidden Markov models (Merialdo, 1994; Brants, 1996, 2000), decision trees (Schmid, 1994; Magerman, 1995), maximum entropy Markov models (MEMMs) (Ratnaparkhi, 1996) and Conditional Random Fields (CRFs) (Lafferty et al., 2001; Constant and Tellier, 2012). Recently, neural approaches have reached very competitive accuracy levels, improving over the state of the art in a number of settings (Plank et al., 2016).

As a complement to annotated training corpora, external lexicons can be a valuable source of information. First, morphosyntactic lexicons provide a large inventory of (word, PoS) pairs. Such lexical information can be used in the form of constraints at tagging time (Kim et al., 1999; Hajič, 2000) or during the training process as additional features combined with standard features extracted from the training corpus (Chrupała et al., 2008; Goldberg et al., 2009; Denis and Sagot, 2012).

Second, lexical information encoded in vector representations, known as word embeddings, have emerged more recently (Bengio et al., 2003; Collobert and Weston, 2008; Chrupała, 2013; Ling et al., 2015; Ballesteros et al., 2015; Müller and Schütze, 2015). Such representations, often extracted from large amounts of raw text, have proved very useful for numerous tasks including PoS tagging, in particular when used in recurrent neural networks (RNNs) and more specifically in mono- or bi-directional, word-level or character-level long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997; Ling et al., 2015; Ballesteros et al., 2015; Plank et al., 2016).

Character-level embeddings are of particular interest for PoS tagging as they generate vector representations that result from the internal character-level make-up of each word. It can generalise over relevant sub-parts such as prefixes or suffixes, thus directly addressing the problem of unknown words. However, unknown words do not always follow such generalisations. In such cases, character-level models cannot bring any advantage. This is a difference with external lexicons, which provides information about any word it contains, yet without any quantitative distinction between relevant and less relevant information.

Therefore, a comparative assessment of the ad-

Proceedings of the 15th International Conference on Parsing Technologies, pages 25–31,
Pisa, Italy; September 20–22, 2017.

vantages of using character-level embeddings and external lexical information is an interesting idea to follow. However, the inclusion of morphosyntactic information from lexicons into neural PoS tagging architecture, as a replacement or complement to character-based or pre-computed word embeddings, remains to be investigated. In this paper, we describe how such an inclusion can be achieved and show, based on experiments using the Universal Dependencies corpora (version 1.3), that it leads to significant improvements over Plank et al.'s (2016) state-of-the-art results.

2 Baseline bi-LSTM tagger

As shown by Plank et al. (2016), state-of-the-art performance can be achieved using a bi-LSTM architecture fed with word representations. Optimal performance is achieved representing words using the concatenation of (i) a word vector $\vec{w}$ built using a word embedding layer, called its *word embedding*, and (ii) a representation $\vec{c}$ of the word's characters, called its *character-based embedding* built using a character-level bi-LSTM, which is trained jointly with the word-level layers. Further improvements can be obtained on most but not all languages by initialising the word embedding layer with pre-computed word embeddings. We refer to Plank et al. (2016) for further details.

3 Integrating lexical information

We extend this bi-LSTM architecture with an additional input layer that contains token-wise features obtained from a lexicon. The input vector $\vec{l}$ for a given word is an n-hot vector where each active value corresponds to one of the possible labels in the lexicon. For instance, the English word *house*, which is both a singular noun and a verb in its base form, will be associated to a 2-hot input vector. Words that are not in the lexicon are represented in the form of a zero vector. Note there is no need for the morphosyntactic features to be harmonized with the tagset to predict.

Figure 1 shows how the output of this input layer is concatenated to that of the two baseline input layers, i.e. the word embedding $\vec{w}$ and (if enabled) the character-based embedding $\vec{c}$. The result of this concatenation feeds the bi-LSTM layer.

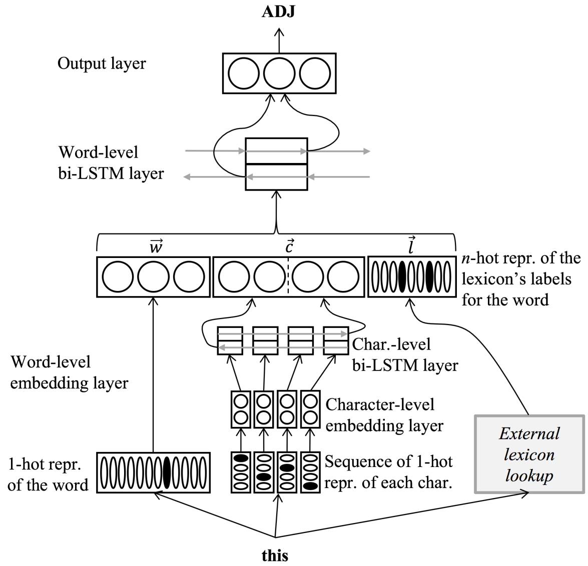

Figure 1: Schema of our extension of Plank et al.'s (2016) bi-LSTM tagging architecture for integrating external morphosyntactic lexical information. This schema concerns a single word, here "this." Connections of the word-level LSTM cell to its counterparts for the preceeding and following word are represented with grey arrows.

4 Data

We use the Universal Dependencies (UD) datasets for our experiments. In order to facilitate comparison with Plank et al.'s (2016), we performed our experiments on the version 1.3 of UD (Nivre et al., 2016).

Lexicons Our sources of lexical information we used are twofold. The first one is the Apertium[2] and the Giellatekno[3] projects. We used Apertium morphological lexicons whenever available. For other languages, we downloaded the corresponding monolingual part of OPUS's OpenSubtitles2016 corpus, tokenised it, extracted the 1 million most frequent tokens, and retrieved all their morphological analyses by the corresponding morphological analyser provided by Apertium (or, failing that, Giellatekno). All these analyses were then gathered in the form of a lexicon. In a second step, we converted all lexicons obtained using manually crafted rules, so that each lexical entry contains a (inflected) wordform, a lemma, a Universal PoS,[4] and morphological features from the Universal Features.[5] We then created two variants of the lexicons obtained: a *coarse* variant in which labels are Universal PoS, and a *full* variant

[2]https://svn.code.sf.net/p/apertium/svn/languages
[3]https://victorio.uit.no/langtech/trunk/langs
[4]http://universaldependencies.org/u/pos/all.html
[5]http://universaldependencies.org/u/feat/all.html

	Name	#entries ($\times 10^3$)	#tags	TTR	PG
ar	Apertium	651	15	yes	
bg	Multext-East	53	12	0.18	yes
ca	Apertium	379	13	0.06	yes
cs	Apertium	1,875	15	0.10	yes
da	Apertium	683	15	0.19	yes
de	DeLex	465	52	0.18	yes
el	Apertium	47	12	0.20	yes
en	Apertium	127	12	0.09	yes
es	Le*ffe*	756	34	0.12	yes
et	GiellateknoMA	44	12	0.23	yes
eu	$\text{Apertium}_{\text{full}}$	53	14	0.22	yes
fa	PerLex	512	37	0.10	yes
fi	GiellateknoMA	228	13	0.29	yes
fr	Le*fff*	539	25	0.11	yes
ga	inmdb	114	32	0.26	yes
gl	Apertium	241	12	0.12	no
grc	Diogenes	1,314	18	0.20	no
he	Apertium	268	16	0.12	yes
hi	Apertium	159	14	0.05	yes
hr	HML	1,361	22	0.21	yes
id	$\text{Apertium}_{\text{full}}$	12	38	0.18	no
it	Apertium	278	14	0.10	yes
kk	ApertiumMA	434	16	0.48	no
la	Diogenes	562	16	0.31	no
lv	Apertium	314	14	0.33	no
nl	Alpino lexicon	81	65	0.14	yes
no	Apertium	2,470	13	0.11	yes
pl	Apertium	1,316	15	0.31	yes
pt	Apertium	159	155	0.13	yes
ro	Multext-East	378	14	0.18	no
ru	Apertium	4,401	16	0.32	no
sl	Apertium	654	14	0.24	yes
sv	Saldo	1,215	214	0.17	yes
tr	ApertiumMA	417	14	0.32	no
zh	Apertium	8	13	0.16	no

Table 1: Dataset information. Best per-language lexicon along with its size and number of tags over the UD1.3 corpora. "MA" stands for morphological-analyser-based lexicon. Lexicons based on Apertium and Giellatekno data are in their *coarse* version unless *full* is indicated. Other lexicons have been adapted from available resources.[1] We also provide the type-token ratio of the corpus (TTR) and whether there were available Polyglot embeddings (PG) to initialize $\vec{w}$.

Lang	Coverage (%)		
	OOTC	OOTC, in Lex.	OOLex
ar	8,0	1,0	55,0
bg	12,3	4,6	32,6
ca	4,9	2,5	20,5
cs	7,0	2,9	31,7
da	15,6	7,3	29,0
de	11,9	5,3	15,1
el	13,4	2,0	52,7
en	9,1	2,6	26,1
es	7,3	3,5	11,3
et	16,9	1,4	48,9
eu	17,8	2,3	57,7
fa	8,2	2,9	31,0
fi	24,4	4,0	46,0
fr	5,7	3,0	9,9
ga	22,8	7,2	66,5
gl	9,9	5,9	14,9
grc	17,9	13,6	57,6
he	10,9	5,1	28,4
hi	4,6	1,6	17,4
hr	20,9	15,1	16,5
id	13,8	2,4	38,3
it	5,7	3,4	21,4
kk	40,5	30,7	23,0
la	26,4	23,4	3,5
lv	36,3	16,9	42,6
nl	18,8	4,4	27,6
no	11,2	4,0	33,0
pl	23,1	9,1	38,9
pt	8,6	3,0	29,2
ro	12,1	6,8	33,1
ru	26,0	15,5	38,7
sl	19,9	11,1	28,7
sv	14,9	10,4	10,4
tr	24,8	13,3	25,6
zh	12,5	0,5	66,5

Table 2: Coverage of the training set and of the best lexicon on the test set for each dataset of the UD 1.3 corpora. "OOTC" stands for "out of training corpus" and OOLex for "out of (external) lexicon". The "OOTC, in Lex." column displays the percentage of words that are not in the training corpus but are covered by the lexicon. Best improvements are expected for these words.

in which labels are the concatenation of the Universal PoS and Universal Features.

We also took advantage of other existing lexicons. For space reasons, we are not able to describe here the language-specific transformations we applied to some of these lexicons. See Table 1 and its caption for more information. We determine the best performing lexicon for each language based on tagging accuracy on the development set. In the remainder of this paper, all information about the lexicons (Table 1) and accuracy results are restricted to these best performing lexicons.

Coverage information on the test sets for both the training data and the best external lexicon for each dataset is provided in Table 2.

Pre-computed embeddings Whenever available and following Plank et al. (2016), we performed experiments using Polyglot pre-computed embeddings (Al-Rfou et al., 2013). Languages for which Polyglot embeddings are available are indicated in Table 1.

We trained our tagger with and without character-based embeddings, and with or without Polyglot-based initialisation (when available), both without lexical information and with lexicon information from all available lexicons, resulting in 4 to 12 training configurations.

Language	Baseline (no lexicon)			With best lexicon (selected on dev, cf. Tab. 1)			Gain when using best lexicon		
	$\vec{w}$	$\vec{w}+\vec{c}$	$\vec{w}_P+\vec{c}$	$\vec{w}+\vec{l}$	$\vec{w}+\vec{c}+\vec{l}$	$\vec{w}_P+\vec{c}+\vec{l}$	$\vec{w}(+\vec{l})$	$\vec{w}+\vec{c}(+\vec{l})$	$\vec{w}_P+\vec{c}(+\vec{l})$
Arabic (ar)	93.90	95.99	96.20	94.58	96.05	96.22	+0.68	+0.06	+0.02
Bulgarian (bg)	94.50	98.11	97.62	96.29	98.30	97.86	+1.79	+0.18	+0.24
Catalan (ca)	96.14	98.03	98.17	97.58	98.21	98.26	+1.44	+0.18	+0.09
Czech (cs)	95.93	98.03	98.10	96.74	98.46	98.41	+0.81	+0.43	+0.31
Danish (da)	90.16	95.41	95.62	94.20	96.24	96.14	+4.04	+0.83	+0.53
German (de)	87.94	92.64	92.96	91.52	93.08	93.18	+3.58	+0.44	+0.23
Greek (el)	95.62	97.76	98.22	96.03	97.67	98.17	+0.41	–0.09	–0.05
English (en)	91.12	94.38	94.56	92.97	94.63	94.70	+1.85	+0.25	+0.14
Spanish (es)	93.10	94.96	95.27	94.62	94.84	95.07	+1.52	–0.11	–0.20
Estonian (et)	90.73	96.10	96.40	90.07	96.14	96.66	–0.65	+0.04	+0.26
Basque (eu)	88.54	94.34	95.07	88.52	94.78	95.03	–0.02	+0.44	–0.04
Persian (fa)	95.57	96.39	97.35	96.22	97.09	97.35	+0.65	+0.71	+0.00
Finnish (fi)	87.26	94.84	95.12	88.67	94.87	95.13	+1.40	+0.03	+0.01
French (fr)	94.30	95.97	96.32	95.92	96.71	96.28	+1.62	+0.74	–0.04
Irish (ga)	86.94	89.87	91.91	88.88	91.18	91.76	+1.94	+1.31	–0.16
Galician (gl)	94.78	96.94	—	95.72	97.18	—	+0.94	+0.24	—
Ancient Greek (grc)	88.69	94.40	—	89.76	93.75	—	+1.07	-0.65	—
Hebrew (he)	92.82	95.05	96.57	94.11	95.53	96.76	+1.29	+0.48	+0.19
Hindi (hi)	95.55	96.22	95.93	96.22	96.50	96.95	+0.67	+0.28	+1.02
Croatian (hr)	86.62	95.01	95.93	93.53	96.29	96.34	+6.91	+1.28	+0.41
Indonesian (id)	89.07	92.78	93.27	91.17	92.79	92.89	+2.11	+0.02	–0.38
Italian (it)	95.29	97.48	97.77	97.54	97.81	97.88	+2.26	+0.33	+0.11
Kazakh (kk)	72.74	76.32	—	82.28	82.79	—	+9.54	+6.47	—
Latin (la)	85.18	92.18	—	90.63	93.29	—	+5.44	+1.12	—
Latvian (lv)	78.22	89.39	—	83.56	91.07	—	+5.35	+1.68	—
Dutch (nl)	84.91	89.97	87.80	85.20	90.69	89.85	+0.29	+0.72	+2.05
Norwegian (no)	93.65	97.50	97.90	95.80	97.72	97.96	+2.15	+0.22	+0.07
Polish (pl)	87.99	96.21	96.90	90.81	96.40	97.02	+2.83	+0.18	+0.13
Portuguese (pt)	93.61	97.00	97.27	94.76	96.79	97.11	+1.15	–0.21	–0.16
Romanian (ro)	92.63	95.76	—	94.49	96.26	—	+1.86	+0.51	—
Russian (ru)	84.72	95.73	—	93.50	96.32	—	+8.79	+0.60	—
Slovene (sl)	83.96	97.30	95.27	94.07	97.74	95.44	10.11	+0.44	+0.17
Swedish (sv)	92.06	96.26	96.56	95.61	97.03	97.00	+3.55	+0.77	+0.44
Turkish (tr)	87.02	93.98	—	90.03	93.90	—	+3.01	–0.08	—
Chinese (zh)	89.17	92.99	—	89.29	93.04	—	+0.12	+0.05	—
Macro-avg.	90.01	94.61	—	92.60	95.18	—	+2.59	+0.57	—
Macro-avg. w/embed	91.43	95.52	95.77	93.52	95.91	95.98	+2.09	+0.38	+0.21

Table 3: Overall results. PoS accuracy scores are given for each language in the baseline configuration (the same as Plank et al., 2016) and in the lexicon-enabled configuration. For each configuration, scores are given when using word embeddings only ($\vec{w}$), word and character-based embeddings ($\vec{w}+\vec{c}$), and word and character-based embeddings with initialisation of word embeddings with Polyglot vectors ($\vec{w}_P+\vec{c}$). The last columns show the difference between lexicon-enabled and baseline configurations.

5 Experimental setup

We use as a baseline the state-of-the-art bi-LSTM PoS tagger `bilty`, a freely available[6] and "significantly refactored version of the code originally used" by Plank et al. (2016). We use its standard configuration, with one bi-LSTM layer, character-based embeddings size of 100, word embedding size of 64 (same as Polyglot embeddings), no multitask learning,[7] and 20 iterations for training.

We extended `bilty` for enabling integration of lexical morphosyntactic information, in the way described in the previous section.

For each lexicon-related configuration, we trained three variants of the tagger: (i) a variant without using character-based embeddings and standard (zero) initialisation of word embeddings before training, (ii) a variant with character-based embeddings and standard initialisation of word embeddings, and (iii) when Polyglot embeddings are available for the language at hand, a variant with character-based embeddings and initialisation of the word embeddings with the Polyglot embeddings. This is deliberately similar to Plank et al.'s (2016) experimental setup, in order to facilitate the comparison of results.[8]

[5]Bouma et al., 2000; Oliver and Tadić, 2004; Heslin, 2007; Borin et al., 2008; Molinero et al., 2009; Sagot, 2010; Erjavec, 2010; Sagot and Walther, 2010; Měchura, 2014; Sagot, 2014.

[6]https://github.com/bplank/bilstm-aux

[7]Plank et al.'s (2016) secondary task—predicting the frequency class of each word—results in better OOV scores but virtually identical overall scores when averaged over all tested languages/corpora.

[8]Note that we discarded alternative UD 1.3 corpora (e.g. `nl_lassysmall` vs. `nl`), as well as corpora for languages for which we had neither a lexicon nor Polyglot embeddings (Old Church Slavonic, Hungarian, Gothic, Tamil).

lang	$w_{(\vec{P})} + \vec{c} + \vec{l}$		Δ w.r.t. $w_{(\vec{P})} + \vec{c}$	
	OOTC	OOTC in Lex.	OOTC	OOTC in Lex.
ar	82,09	94,78	-0,53	-0,51
bg	92,79	96,84	+4,67	+0,98
ca	94,21	98,38	+0,31	-0,11
cs	90,84	96,82	+5,21	+0,57
da	88,54	95,03	+3,17	+0,70
de	86,05	87,00	+3,32	+0,41
el	89,22	96,52	-1,97	-0,90
en	78,23	89,31	+3,89	+1,02
es	76,34	79,33	-1,21	-1,12
et	88,24	94,80	-1,62	-0,70
eu	82,02	93,26	-0,09	-0,41
fa	84,94	95,34	-1,22	-0,76
fi	85,31	92,03	-0,76	-0,95
fr	85,50	86,35	+2,25	+0,43
ga	77,43	89,09	-0,34	-1,77
gl	85,20	91,21	+21,73	+5,60
grc	83,71	94,40	+25,16	+2,00
he	81,36	92,25	-5,81	-2,61
hi	78,91	93,84	-4,22	-0,78
hr	90,74	88,66	+1,50	+0,44
id	86,07	90,72	-1,29	-0,55
it	89,15	96,46	+1,12	-0,43
kk	76,89	52,59	+23,53	-2,96
la	84,51	88,89	+28,95	+10,53
lv	80,98	83,64	+35,13	+15,83
nl	69,49	78,60	+12,75	+8,19
no	92,44	96,97	-0,24	-0,48
pl	90,48	93,95	-2,65	-2,04
pt	88,13	95,69	+0,19	-0,60
ro	88,39	95,47	+23,18	+3,71
ru	90,49	93,80	+40,87	+13,05
sl	93,31	95,77	+11,56	+4,41
sv	92,43	93,31	+3,88	-0,47
tr	85,33	87,33	+26,68	+9,13
zh	78,30	92,08	+24,97	+5,07
Macro avg.	85,37	90,87	+8,06	+1,83

Table 4: Accuracy of the best system using a lexicon for words out of the training corpus (OOTC), and for words out of the training corpus that are present in the lexicon (OOTC in Lex.), as well as difference between the best system and the baseline without lexicon for these two subsets of words.

6 Results

Our results show that using lexical information as an additional input layer to a bi-LSTM PoS tagger results in consistent improvements over 35 corpora. The improvement holds for all configurations on almost all corpora. As expected, the greatest improvements are obtained without character-based embeddings, with a macro-averaged improvement of +2.56, versus +0.57 points when also using character-based embeddings. When also using pre-computed embeddings, improvements are only slightly lower. External lexical information is useful as it covers both words with an irregular morphology and words not present in the training data.

The improvements are particularly high for the smaller datasets; in the $\vec{w} + \vec{c}$ setup, the three languages with the highest improvements when using a lexicon are those with smallest datasets.

Table 4 shows the accuracy of the best system, compared with the baseline, for words not in the training data (OOTC), and for whose that are present in the lexicon but not in the training data (OOTC in Lex).

While lexicon coverage is an important, it is not the only factor. we observe the improvements are much larger for the smaller datasets like Kazakh (kk) or Russian (ru). However, the improvement is smaller for words that are not in the training data but are nevertheless present in the lexicon, which indicates that the contribution of the lexicon features to PoS prediction is not limited to the words that are covered by the lexicon but spreads throught the contexts by means of the bi-LSTM architecture. Moreover, we argue that the presence of the lexicon features aids compensate for character embeddings fit on smaller datasets, which are not necessarily more trustworthy.

7 Conclusion

Our work shows that word embeddings and external lexical information are complementary sources of morphological information, which both improve the accuracy of a state-of-the-art neural part-of-speech tagger. It also confirms that both lexical information and character-based embeddings capture morphological information and help part-of-speech tagging, especially for unknown words.

Interestingly, we also observe improvements when using external lexical information together with character-based embeddings, and even when initialising with pre-computed word embeddings. This shows that the use of character-based embeddings is not sufficient for addressing the problem of out-of-vocabulary words.

Further work includes using lexicons to tag finer-grained tag inventories, as well as a more thorough analysis on the relation between lexicon and training data properties.

Another natural follow-up to the work presented here would be to examine the interplay between lexical features and more complex neural architectures, for instance by using more than one bi-LSTM layer, or by embedding the n-hot lexicon-based vector before concatenating it to the word- and character-based embeddings.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proc. of the Seventeenth Conf. on Computational Natural Language Learning*. Sofia, Bulgaria, pages 183–192.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 349–359.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3(1):1137–1155.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for swedish language technology. In *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, Uppsala University, Uppsala, Sweden, pages 21–32.

Gosse Bouma, Gertjan van Noord, and Rob Malouf. 2000. Alpino: Wide-coverage computational analysis of dutch. In *Computational Linguistics in the Netherlands 2000, Selected Papers from the Eleventh CLIN Meeting, Tilburg, November 3, 2000*. pages 45–59.

Thorsten Brants. 1996. Estimating markov model structures. In *Proc. of the 4th Conf. on Spoken Language Processing (ICSLP-96)*. pages 893–896.

Thorsten Brants. 2000. TnT: A Statistical Part-of-speech Tagger. In *Proc. of the Sixth Conf. on Applied Natural Language Processing*. Seattle, Washington, USA, pages 224–231.

Grzegorz Chrupała. 2013. Text segmentation with character-level text embeddings. In *Proc. of the ICML Workshop on Deep Learning for Audio, Speech and Lang. Processing*. Atlanta, Georgia, USA.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with morfette. In *Proc. of the 6th Language Resource and Evaluation Conf.*. Marrakech, Morocco.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of the 25th International Conf. on Machine Learning*. Helsinki, Finland, pages 160–167.

Matthieu Constant and Isabelle Tellier. 2012. Evaluating the Impact of External Lexical Resources into a CRF-based Multiword Segmenter and Part-of-Speech Tagger. In *Proc. of LREC'12*. Istanbul, Turkey, pages 646–650.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 46(4):721–736.

Tomaž Erjavec. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of LREC 2010*. Valletta, Malta.

Y. Goldberg, R. Tsarfaty, M. Adler, and M. Elhadad. 2009. Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and em-hmm-based lexical probabilities. In *Proc. of the 12th Conf. of the European Chapter of the ACL*. pages 327–335.

Jan Hajič. 2000. Morphological Tagging: Data vs. Dictionaries. In *Proc. of ANLP'00*. Seattle, Washington, USA, pages 94–101.

Peter J. Heslin. 2007. Diogenes, version 3.1. `http://www.dur.ac.uk/p.j.heslin/Software/Diogenes/`.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neur. Comp.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

J.-D. Kim, S.-Z. Lee, and H.-C. Rim. 1999. HMM Specialization with Selective Lexicalization. In *Proc. of the join SIGDAT Conf. on Empirical Methods in Natural Lang. Processing and Very Large Corpora*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*. pages 282–289.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proc. of the 2015 Conf. on Empirical Methods in Natural Lang. Processing*. Lisbon, Portugal, pages 1520–1530.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proc. of the 33rd Annual Meeting on ACL*. Cambridge, Mass., USA, pages 276–283.

Bernard Merialdo. 1994. Tagging English Text with a Probabilistic Model. *Computational Linguistics* 20(2):155–171.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The leffe. In *Proc. of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*. Borovets, Bulgaria.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proc. of the 2015 Conf. of the North American Chapter of the ACL: Human Language Technologies*. Denver, Colorado, USA.

Michal Boleslav Měchura. 2014. Irish National Morphology Database: A High-Accuracy Open-Source Dataset of Irish Words. In *Proc. of the Celtic Language Technology Workshop at CoLing*. Dublin, Ireland.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Çar Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Simon Krek, Veronika Laippala, Lucia Lam, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Loganathan Ramasamy, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jing Xian Wang, Jonathan North Washington, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-1699.

Antoni Oliver and Marko Tadić. 2004. Enlarging the Croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Proc. of LREC 2004*. Lisbon, Portugal, pages 1259–1262.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proc. of the 54th Annual Meeting of the ACL*. Berlin, Germany.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. of International Conf. on Empirical Methods in Natural Language Processing*. pages 133–142.

Benoît Sagot. 2010. The le*fff*, a freely available, accurate and large-coverage lexicon for french. In *Proc. of LREC 2010*. Valletta, Malta.

Benoît Sagot. 2014. DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German. In *Language Resources and Evaluation Conf.*. Reykjavik, Iceland.

Benoît Sagot and Géraldine Walther. 2010. A morphological lexicon for the Persian language. In *Proc. of LREC 2010*. Valletta, Malta.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conf. on New Methods in Language Processing*. Manchester, UK.