Do LSTMs really work so well for PoS tagging? – A replication study

Tobias Horsmann and Torsten Zesch
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany
{tobias.horsmann,torsten.zesch}@uni-due.de

Abstract

A recent study by Plank et al. (2016) found that LSTM-based PoS taggers considerably improve over the current state-of-the-art when evaluated on the corpora of the Universal Dependencies project that use a *coarse-grained* tagset. We replicate this study using a fresh collection of 27 corpora of 21 languages that are annotated with *fine-grained* tagsets of varying size. Our replication confirms the result in general, and we additionally find that the advantage of LSTMs is even bigger for larger tagsets. However, we also find that for the very large tagsets of morphologically rich languages, hand-crafted morphological lexicons are still necessary to reach state-of-the-art performance.

1 Introduction

Part-of-Speech (PoS) tagging is an important processing step for many NLP applications. When researchers want to use a PoS tagger, they would ideally choose an off-the-shelf PoS tagger which is optimized for a specific language. If a suited tagger is not available two options remain: a) implementation of your own tagger, which requires technical knowledge and experience, or b) using an existing tagger and hope that the resulting model will be sufficiently accurate. One can assume that many taggers fit more languages than the one for which they have been constructed originally. Ideally, researchers should be able to fall back to a well-evaluated language-independent tagger if no reference implementation for a language is available.

A recent study by Plank et al. (2016) evaluated an LSTM PoS tagger and compared the results to Conditional Random Fields (CRF) (Lafferty et al., 2001) and Hidden-Markov (HMM) implementations on corpora of various languages. Their evaluation concludes that the LSTM tagger reaches better results than the CRF and HMM tagger. The evaluation corpora were all annotated with a *coarse-grained* tagset with 17 tags. Thus, this LSTM tagger seems to be a well-performing, language-independent choice for learning models on coarse-grained tagsets. While for many tasks a coarse-grained tagset might be sufficient some tasks require more fine-grained tagsets.

We, thus, consider it worthwhile to explore if the results are reproducible using corpora with fine-grained tagsets. We use the LSTM tagger provided by Plank et al. (2016) and compare the results likewise to CRF and an off-the-shelf HMM tagger implementation. We compile a fresh set of 27 corpora of 21 languages which uses the commonly used *fine-grained* tagset of the respective language. We suggest these corpora as evaluation set for tasks which require fine-grained PoS tags, as all corpora are freely available for research purposes. Our intention is to replicate the findings of Plank et al. (2016), which have been achieved on a coarse-grained tagset and investigate if they transfer to fine-grained tagsets.

2 PoS Tagger Paradigms

We distinguish two PoS tagger paradigms, which can be used to implement a tagger: The first one is *Feature Engineering*, in which a classifier learns a mapping from human-defined features to a PoS tag. Defining good features is often a non-trivial task, which furthermore requires a lot of experience. For instance a suffix feature which checks a word-ending for “ing” is highly discriminative for English gerunds, but might not provide any useful information for other languages. The details of the feature implementation might render a

Group	Corpus Id	Source	Tokens (10^3)	# Tags	Annotation	Reference
Germanic	Danish	Copenhagen DTB	255	36	manual	(Buch-Kromann and Korzen, 2010)
	Dutch	Alpino	200	20	manual	(Bouma et al., 2000)
	English	Brown	1,100	180	manual	(Nelson Francis and Kuçera, 1964)
	German-1	Hamburg DTB	4,800	54	manual	(Brants et al., 2004)
	German-2	Tiger	880	54	manual	(Telljohann et al., 2004)
	German-3	Tüba-D/Z	1,500	54	manual	(Foth et al., 2014)
	Icelandic	Mim	1,000	703	auto	(Helgadóttir et al., 2012)
	Norwegian	Norwegian DTB	1,300	19	manual	(Solberg et al., 2014)
	Swedish-1	Talbanken	96	25	manual	(Einarsson, 1976)
	Swedish-2	Stockholm-Umea	1,100	153	manual	(Ejerhed and Källgren, 1997)
Romanic	Braz.Portuguese	MAC-Morpho	1,000	82	manual	(Aluísio et al., 2003)
	French-1	Multitag	370	992	manual	(Paroubek, 2000)
	French-2	Sequoia	200	29	manual	(Candito et al., 2014)
	Italian	Turin Parallel	80	15	auto	(Bosco et al., 2012)
	Spanish	IULA DTB	550	241	manual	(Marimon et al., 2014)
Slavic	Croatian-1	Croatian DTB	200	692	manual	(Željko Agić and Ljubešić, 2014)
	Croatian-2	Hr500k	500	769	manual	(Ljubešić et al., 2016)
	Czech	Prague DTB	2,000	1,574	manual	(Bejček et al., 2013)
	Polish	Polish National Corpus	1,000	27	manual	(Przepiórkowski et al., 2008)
	Russian	Russian Open Corpus	1,700	22	manual	(Bocharov et al., 2013)
	Slovak	MULTEXT-East	84	956	manual	(Erjavec, 2010)
	Slovene-1	IJS-ELAN	540	1,181	auto	(Erjavec, 2002)
	Slovene-2	SSJ	590	1,304	manual	(Krek et al., 2013)
Others	Afrikaans	AfriBooms	50	12	manual	(Augustinus et al., 2016)
	Finnish	FinnTreebank	170	1573	manual	(Voutilainen, 2011)
	Hebrew	HaAretz Corpus	11,000	22	auto	(Itai and Wintner, 2008)
	Hungarian	The Szeged Treebank	1,200	1,085	manual	(Csendes et al., 2005)

Table 1: Corpora used in our experiments

tagger unsuited for learning models for other languages or tagsets. We will, thus, experiment with features and their configurations, and investigate how well they perform in combination for learning fine-grained tagsets of various languages. We implement those experiments using CRF which are frequently used for PoS tagging (Remus et al., 2016; Ljubešić et al., 2016).

The second paradigm is *Architecture Engineering*, which relies on methods to learn the input representation by themselves. The challenge lies in finding an architecture that supports this self-learning process. Most recent representatives of this paradigm are neural networks of which we use the LSTM tagger provided by Plank et al. (2016).

In our experiments, we will focus on how to provide word- and character-level information to the classifiers as these two types of information are most relevant and most frequently used for training PoS tagger models. Furthermore, we will evaluate the performance on Out-Of-Vocabulary (OOV) words to learn if the taggers generalize to unseen words.

To provide a reference value to a well-known PoS tagger, we will compare all results to the HMM-based HunPos (Halácsy et al., 2007) tagger, which is a freely available re-implementation of the TNT tagger (Brants, 2000). HunPos has been used before for training models of various languages and tagsets (Seraji, 2011; Attardi et al., 2010; Hládek et al., 2012) which is why we consider this tagger to be a suitable baseline.

3 Evaluation Corpora Dataset

Table 1 shows the fine-grained annotated corpora we collected by screening the literature. We do not claim that this list is complete, but the provided corpora are all reasonably easy to access and can be freely used for research purposes.

Selection To ensure reproducibility, we preferably selected corpora which are directly available via the Internet except *German-3, Hungarian* and *Swedish-2*. We intentionally exclude languages such as Chinese or Japanese, which do not provide whitespace delimiters to mark word boundaries. Tagging those languages requires a morpho-

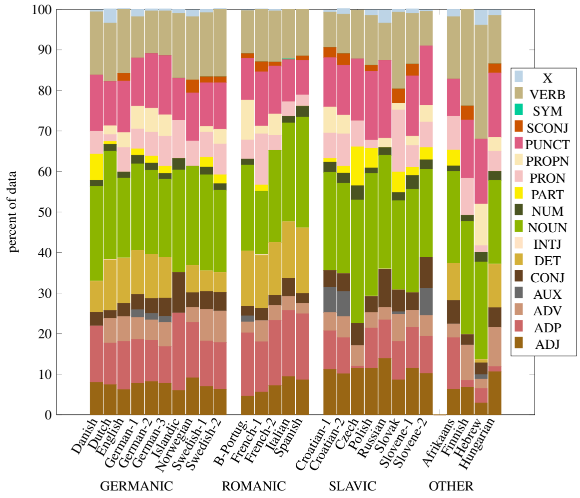

Figure 1: Coarse-grained PoS tag distribution of corpora by language group

logical analysis which is a different task than the tagging task on which we are focusing here. Most corpora are manually annotated or were at least human-verified. There are four exceptions which we decided to add anyway to increase the number of languages represented in our setup. The tagset granularity of the corpora ranges from coarse (12 tags) to morphologically fine (1574 tags) to evaluate all taggers on various stages of granularity.

Language & Corpora Diversity We analyzed the distribution of PoS tags in the corpora by mapping all tags to the 17 coarse-grained PoS tags of the Universal Dependencies (UD) project (Nivre et al., 2015) in Figure 1. The mappings to the UD tagset have been manually created. The partly large differences between the syntactical classes help to better understand the challenge in construction a tagger that is suited for all those languages. For instance, Germanic and Romanic languages have a lot of determiners while they do not occur at all in Slavic languages.

Corpus Size & Tagset The corpora have varying sizes which makes a direct comparison between corpora difficult. To run our experiments under fully controlled conditions, we extract a randomized sub-sample of sentences from each corpus, which accounts for 50k tokens, and run all our experiments with 10fold cross-validation (CV).[1] Results reported use the fine-grained tagset of the respective corpus.

[1] While randomization prohibits exact reproducibility, it is no barrier to the more interesting replicability. It is also less prone to continued overfitting on the known test set.

We deliberately do not use the corpora from the UD Treebank project in order to provide results on a fresh dataset. Additionally, UD uses a coarse-grained tagset for all its corpora. While this granularity is sufficient for many tasks, linguistic analysis often requires more fine-grained tagsets, and it is not clear whether results achieved on coarse-grained tagsets transfer well to more fine-grained tagsets. The collected corpora, thus, also represent an alternative dataset, which we suggest in case the UD tagset is too coarse-grained.

4 CRF Experiments

We reviewed the recent literature to determine the most commonly used features for training PoS taggers. As re-occurring features, we found word ngrams, fixed character sequences focusing on either pre-, in-, or suffixes of words and word distributional knowledge for PoS taggers of various languages (Brants, 2000; Horsmann and Zesch, 2016; Ljubešić et al., 2016). Word- and character-ngrams have been used with various parametrizations depending on the language and there is no agreement which parameters are most advisable. We will, hence, run a series of parameter-search experiments over the word- and character-ngram parametrization to determine a configuration applicable to all languages. For this, we evaluate all permutations of the subsequently introduced feature configurations with 10fold cross-validation. The objective is to find a configuration that works well on all corpora, languages, and tagsets.

Word Features We experiment with adding the $1,2,3$ words to the right and left of the current word as lower-cased string features.

Character Features Which character-ngram is discriminative for a PoS tag strongly depends on the language. To avoid a language bias, we use a frequency-based approach in which we select the N most frequently occurring character-ngrams of length $1,2,3,4$ from the training dataset. We experiment with the following frequency cut-off values of $N \; \varepsilon \; \{250, 500, 750, 1000\}$ to select only frequent and potentially informative character-ngrams as features. These N features are boolean and are set to 1 if the respective character-ngram occurs in the current word.

Semantic Features We use Brown clustering (Brown et al., 1992) to create word clusters. The

Lang. Group	Corpus Id	Word Ngrams ±1 All	OOV	Top 750 Char Ngrams All	OOV	Clusters All	OOV	Best CRF All	OOV	HunPos All	OOV
Germanic	Danish	90.9	53.3	90.3	69.3	89.5	67.6	96.1	82.4	94.9	74.2
	Dutch	86.5	66.9	85.0	71.7	88.0	77.7	90.7	83.7	89.9	80.6
	English	87.5	45.1	90.3	70.1	89.1	64.0	94.6	80.2	93.8	77.7
	German-1	88.5	62.4	90.3	77.7	90.8	73.7	94.6	84.6	94.4	83.7
	German-2	87.2	60.3	90.9	77.7	90.8	76.1	95.2	87.1	94.9	85.4
	German-3	86.3	58.5	91.7	76.8	91.6	77.6	94.4	85.0	94.4	83.9
	Icelandic	67.5	14.2	76.5	45.1	68.3	28.9	80.9	53.6	79.8	51.9
	Norwegian	92.4	77.1	91.6	80.6	92.8	82.7	96.1	89.7	95.5	86.5
	Swedish-1	91.1	70.6	92.9	82.2	92.3	79.9	96.3	90.3	95.6	85.9
	Swedish-2	78.7	29.7	87.2	67.3	81.4	48.8	91.0	74.6	91.4	77.6
Romanic	B-Portug.	86.9	62.8	87.8	73.6	89.7	76.0	92.8	83.8	93.3	84.2
	French-1	81.9	40.1	85.9	66.5	81.6	58.2	89.2	75.7	88.2	71.8
	French-2	95.4	67.3	93.8	74.5	91.9	79.3	97.7	88.2	97.4	82.4
	Italian	93.3	68.6	91.6	74.8	91.7	75.5	96.4	86.5	95.8	80.8
	Spanish	88.5	45.5	94.5	78.2	88.1	58.8	96.4	83.5	96.6	83.6
Slavic	Croatian-1	69.0	18.6	80.6	56.3	75.2	47.2	84.9	65.4	84.7	66.7
	Croatian-2	66.3	15.9	78.5	54.4	73.5	44.8	83.4	63.9	82.6	63.9
	Czech	64.1	14.4	79.2	56.0	75.2	39.2	83.1	62.9	81.7	60.9
	Polish	82.9	58.1	92.5	86.9	86.5	72.5	95.5	91.5	93.6	85.4
	Russian	83.7	53.7	93.0	83.5	88.2	70.9	95.5	87.5	94.6	83.6
	Slovak	67.7	14.9	80.5	57.8	65.6	31.9	83.5	63.8	82.9	61.6
	Slovene-1	72.6	17.4	83.5	55.6	72.4	39.4	86.4	62.5	82.6	59.6
	Slovene-2	65.4	12.1	78.2	50.5	73.0	39.0	83.0	59.4	86.2	59.5
Other	Afrikaans	95.7	75.0	95.3	80.3	95.8	81.9	97.8	89.6	97.3	85.5
	Finnish	62.6	10.0	77.1	48.5	67.8	33.8	82.3	56.7	81.3	55.8
	Hebrew	82.3	41.7	81.3	60.9	76.3	53.3	90.5	68.5	90.3	60.1
	Hungarian	72.7	13.9	86.7	63.3	72.0	31.7	89.9	69.6	89.4	69.5

Table 2: Accuracy of CRF taggers (10fold CV)

unlabelled text is obtained from the Leipzig Corpus Collection (Quasthoff et al., 2006), which provides large text quantities crawled from the web for many languages. We use $15 \cdot 10^6$ tokens to create the clusters from the same amount of text for all languages. We provide the cluster ids in substrings of varying length to the classifier (Owoputi et al., 2013).

Results In Figure 2, we show the results of our parameter search experiment. The triangles mark the results of the various feature configurations. The diamond symbol shows the configuration which works best over all corpora. We refer to this best working configuration as *Best CRF* subsequently, it uses a word-context window of 1 word to the left and right and the 750 most frequent character [1..4] grams with additionally adding word clusters. Especially for morphologically-rich languages, the spread is quite large which is caused by the lower number of character-ngrams in those configurations. For corpora such as *Slovene-1*, we see that more accurate configurations exist than *Best CRF* but more importantly, the selected configuration is always among the best working ones.

We show the results of *Best CRF* and the performance of the individual features for each language in Table 2, and compare the results to HunPos, the highest accuracies are highlighted in grey. When evaluating the features separately, the character-ngrams reach the highest accuracy on OOV words. Especially on the Slavic language family the character-ngrams perform much better than using only word-ngrams or clusters. Furthermore, using only character-ngrams is often competitive to using only word-ngrams. Hence, a rather naïve strategy to achieving a decent performance on almost any language is to just use

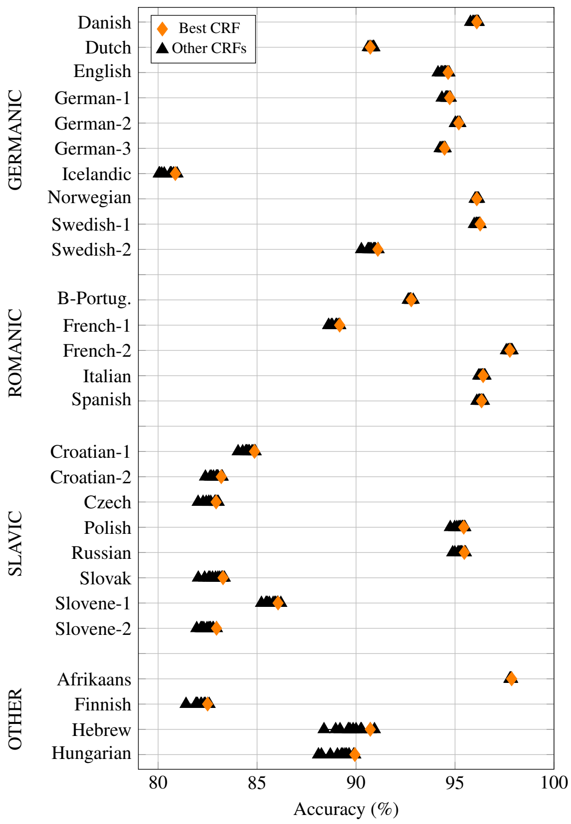

Figure 2: Variance of CRF taggers (10fold CV)

all kinds of character-ngrams. The cluster feature also performs better than the word-ngrams. Considering that we had to limit the amount of data for creating the clusters for comparability, this feature assumedly has more potential when using larger data sizes (Derczynski et al., 2015). The combination of all features in the column *Best CRF* shows that the features address quite different information and add up well, so unsurprisingly, this configuration reaches the overall best accuracies. The difference to HunPos is, with often less than one percent point difference, only small. Off-the-shelf taggers do, hence, not necessary have a disadvantage over constructing an own tagger. In the remainder of this work, we will use the *Best CRF* configuration when discussing CRF tagger results.

5 LSTM Experiments

When using neural networks, the details of how word and character information is provided greatly influences the learning success of the network. We will reproduce network setups which have also been used in Plank et al. (2016) to ensure comparability to the coarse-grained results to which we compare our results:

Word In this setup, we train a network on the word embeddings only and provide them to a bidirectional LSTM. This setup will serve as baseline.

Char The character embeddings of a word are provided to a bidirectional LSTM. The last state of the forward and the backward character LSTM are combined (Ling et al., 2015) and provided to another bidirectional LSTM layer.

Word-Char This architecture is a combination of the previous two architectures. The last state of the character LSTMs is added to the word embedding information before it is provided to the next LSTM layer.

Word-Char+ The architecture by Plank et al. (2016) combines word and character level information and additionally considers the log-frequency of the next word during training. This tagger reported state-of-the-art results and we use the provided reference implementation of this tagger in our setup.

LSTMs have the reputation to require larger amounts of training data. With the 50k tokens we use this is barely fulfilled, however, Plank et al. (2016) find this sensitivity to be less severe and set a corpus size of 60k tokens as lower bound for their coarse-grained tagging experiments. We will come back to this data size issue in Section 7, where we evaluate using all tokens in a corpus (and arriving at the same conclusions as for our 50k token datasets). Furthermore, in many cases only smaller dataset sizes are available, sometimes even less than 50k tokens. It is, thus, important to know if considering neural network taggers makes sense at all (on fine-grained tagsets), thus we will train LSTM models on smaller dataset sizes.

We implement the LSTM taggers in DyNet (Neubig et al., 2017) and use the hyper-parameter settings by Plank et al. (2016), i.e. we train 20 epochs using Statistical-Gradient-Descent with a learning rate of 0.1 and adding Gaussian noise of 0.2 to the embedding layer. We train word embeddings on the data we already used for the *semantic feature* in the CRF experiments by using *fastText* (Bojanowski et al., 2016) . The the character-level embeddings are trained on-the-fly.

Results In Figure 3, we show the results for the LSTM architectures. The *Word-Char+* tagger performs best followed by *Word-Char*, which is not surprising as *Word-Char+* is based on this

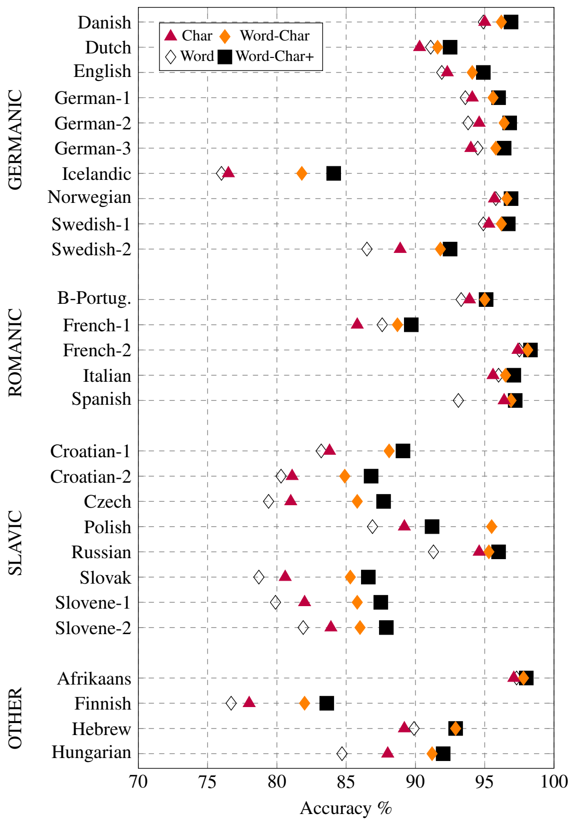

Figure 3: Variance of LSTM taggers (10fold CV)

architecture. For the Germanic and Romanic languages, the accuracy of the various architectures is similar but for Slavic languages, which use much more fine-grained tagsets, the differences are rather large. For instance, the *Char* architecture reaches only small improvements over the *Word* baseline on *Croatian or Czech* while on *Spanish, or Hungarian* the character architecture is clearly better than the baseline. Table 3 shows the detailed results and additionally reports the accuracy values on OOV with best results highlighted in grey. The *Char* architecture is in many cases competitive to the HunPos reference system. This shows that the performance of many off-the-shelf taggers is rather easy to approximate by relying only on character-level information.

The results by the *Char* architecture also explains why the *Word-Char* architecture performs so well although the amount of syntactical information is quite limited with 50k tokens. A large part of the necessary information is already obtained by the character model, which requires a lot less training data than a model on the word level. Thus, the results of Plank et al. (2016) on coarse-tagsets are reproducible for fine-grained tagsets with the *Word-Char* architecture being the essential property to achieving high accuracy.

6 Influence of Tagset Size

A researcher who works with morphologically rich languages will often be interested in additional morphologic details such as case or gender. This drastically complicates the task, as a few hundred instead of a few dozen PoS tag distinctions have to be learned. In this experiment, we will examine the impact of an increasing number of PoS tags on the accuracy of the taggers to provide reference values of how much performance a tagger seems to loose with an increasing tagset size.

Results In Figure 4, we show a comparison of the tagging accuracy in relation to the number of PoS tags. We show the best performing LSTM tagger *Word-Char+*, the *CRF* tagger and *HunPos*. Each data point represents the averaged CV result on one corpus with the respective tagger. We see a certain clustering of the data points for the small tagset sizes, which shows that the taggers tend to perform highly similarly for many languages. This means that the tagset size has a larger effect on the accuracy than the language of the corpus.

For each PoS tagger, a regression trendline is plotted which indicates the average loss in accuracy with an increasing tagset size. For one-hundred additional PoS tags, *Word-Char+* loses 0.35 points in accuracy, while *CRF* and *HunPoS* have a much steeper decay of 0.45 points. Hence, with growing tagset size the tagger choice becomes increasingly more important. Furthermore, the benefit of more sophisticated tagger architectures becomes only apparent on large PoS tagsets.

7 Comparison with Reference Taggers

In this experiment, we compare our results to reference taggers from the literature that are tailored towards certain languages. Our experiments until now were limited to the fixed dataset size that we set at the beginning for comparability. Especially for the morphologically fine-grained tagsets this might have been problematic, as it is doubtful if all PoS tags of a morphological tagset do even occur on 50k tokens. Thus, in order to evaluate the taggers using all available data, we will reproduce setups reported in the literature and compare the performance of the taggers to those results.

This experiment limits the number of comparisons we can make drastically, as we need to have

Lang. Group	Corpus Id	Word All	Word OOV	Char All	Char OOV	Word-Char All	Word-Char OOV	Word-Char+ All	Word-Char+ OOV	HunPos All	HunPos OOV
Germanic	Danish	94.9	72.7	95.0	79.1	96.4	82.5	96.9	83.4	94.9	74.2
	Dutch	91.1	82.3	90.3	83.6	91.6	85.7	92.5	87.1	89.9	80.6
	English	91.9	65.9	92.3	77.4	94.1	79.6	94.9	80.9	93.8	77.7
	German-1	93.6	78.3	94.1	84.5	95.6	87.6	96.0	88.3	94.4	83.7
	German-2	94.5	82.4	94.6	87.1	96.4	90.1	96.8	91.5	94.4	85.4
	German-3	93.8	80.3	94.0	84.9	95.8	88.6	96.4	89.8	94.4	83.9
	Icelandic	76.0	34.8	76.5	49.3	81.8	56.2	84.1	60.6	79.8	51.9
	Norwegian	95.8	86.2	95.7	88.2	96.6	90.3	96.9	90.3	95.5	86.5
	Swedish-1	94.9	81.4	95.3	86.7	96.2	89.0	96.7	89.8	95.6	85.9
	Swedish-2	86.5	54.3	88.9	74.3	91.8	78.5	92.5	80.4	91.4	77.6
Romanic	B-Portug.	93.3	82.4	93.9	87.4	95.0	90.3	95.1	90.8	93.3	84.2
	French-1	87.6	67.0	85.8	72.0	88.7	77.4	89.7	78.7	88.2	71.8
	French-2	97.5	80.4	97.4	83.4	98.1	87.7	98.3	88.7	97.4	82.4
	Italian	96.0	81.3	95.6	84.2	96.5	85.9	97.1	86.9	95.8	80.8
	Spanish	93.1	63.3	96.4	85.5	96.9	86.1	97.2	87.0	96.6	83.6
Slavic	Croatian-1	83.2	55.5	83.8	67.5	88.1	72.8	89.1	75.2	84.7	66.9
	Croatian-2	80.3	52.4	81.1	63.8	84.9	69.1	86.8	72.4	82.6	63.9
	Czech	79.4	49.1	81.0	62.7	85.8	68.7	87.7	72.4	81.7	60.9
	Polish	86.9	73.6	89.2	84.7	95.5	91.2	91.2	88.0	93.6	85.4
	Russian	91.3	73.2	94.6	85.8	95.3	86.9	96.0	88.4	94.6	83.6
	Slovak	78.7	44.9	80.6	65.0	85.3	69.7	86.6	71.4	82.9	61.6
	Slovene-1	81.9	44.5	83.9	61.1	86.0	62.6	87.9	65.7	82.6	59.6
	Slovene-2	79.9	47.9	82.0	63.4	85.8	67.4	87.5	70.1	86.2	59.5
Other	Afrikaans	97.3	82.8	97.1	85.8	97.8	88.4	98.0	90.0	97.3	85.5
	Finnish	76.7	42.7	78.0	57.6	82.0	58.9	83.6	61.2	81.3	55.8
	Hebrew	89.9	60.2	89.2	66.9	92.2	69.7	92.9	72.1	90.3	60.1
	Hungarian	84.7	53.3	88.0	73.1	91.2	76.9	92.0	79.0	89.4	69.5

Table 3: Accuracy of LSTM taggers (10fold CV)

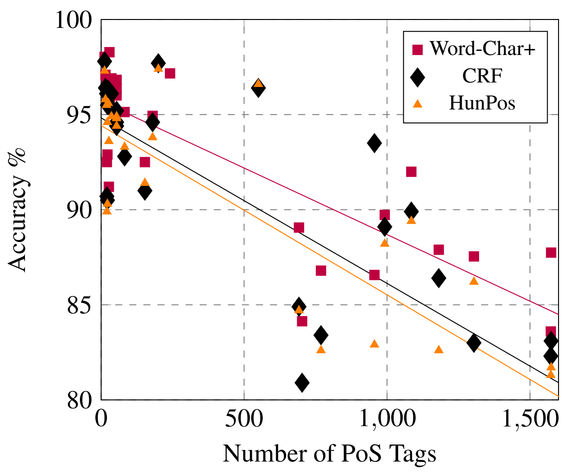

Figure 4: Influence of tagset size on accuracy

the same corpora as used in the literature. We, thus, reproduce for *Czech* the setup by Spoustová et al. (2009) with training on 10^6 and evaluation on $2 \cdot 10^5$ tokens, for *German-2* the setup by Giesbrecht and Evert (2009) and for *Swedish-2* the setup by Östling (2013), which both use 10fold cross-validation over the full corpus size.

Taggers for Slavic languages often make use of additional resources such as morphological dictionaries, which we intentionally do not include to avoid human-crafted resources that are not available for all languages. Thus, we do not expect to reach state-of-the-art performance, but we want to quantify the size of the gap.

Results In Table 4, we show a comparison of our results to the results reported in the literature. On *German-2* and *Swedish-2*, the Word-Char+ tagger is able to reach better results than the reported reference values except for *Czech* which uses a morphologically fine-grained tagset. Thus, language-

Corpus Id	# Tags	Acc (%)	Δ to reference tagger HunPos	CRF	Word-Char+
Czech	1,574	95.9	-4.7	-3.2	-1.5
German-2	54	97.6	-0.1	-0.2	0.9
Swedish-2	153	96.1	0.0	-0.6	0.1

Table 4: Results of reproducing setups in the literature using the *full corpus size*

fitted PoS taggers reach better results than neural networks when training models on corpora with extremely fine-grained PoS tagsets. However, for smaller tagsets sizes the need for using language-fitting is negligible.

8 Conclusion

We replicated a study in which LSTM PoS taggers are compared to CRF and HMM taggers on corpora with a coarse-grained tagset. Our replication focused on whether results reported for coarse-grained tagsets do also hold when training models on fine-grained tagsets. Therefore, we collected a large set of 27 evaluation corpora that are annotated with the commonly used fine-grained tagset of 21 languages. The replication confirmed the superior performance of the LSTM tagger reported by Plank et al. (2016) also on fine-grained tagsets. However, we also found that for smaller tagset sizes the differences between the LSTM, our self-implemented CRF and the HMM tagger are often only small. The advantages of the LSTM tagger over other taggers grow proportionally with the tagsets size of the corpus. On morphologically fine tagsets, even the LSTM tagger fails to reach results reported in the literature when reproducing those setups.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

Željko Agić and Nikola Ljubešić. 2014. The setimes.hr linguistically annotated corpus of croatian. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1724–1727, Reykjavik, Iceland. ELRA.

Sandra Aluísio, Jorge Pelizzoni, Ana Raquel Marchi, Lucélia de Oliveira, Regiana Manenti, and Vanessa Marquiafável. 2003. *An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese*. Faro, Portugal.

Giuseppe Attardi, Stefano Dei Rossi, Giulia Di Pietro, Alessandro Lenci, Simonetta Montemagni, and Maria Simi. 2010. A Resource and Tool for Supersense Tagging of Italian Texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. ELRA.

Liesbeth Augustinus, Peter Dirix, Daniel Van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde, and Gerhard Van Huyssteen. 2016. AfriBooms: An Online Treebank for Afrikaans. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. ELRA.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. Prague Dependency Treebank 3.0.

V. V. Bocharov, S.V. Alexeeva, D.V. Granovsky, E.V. Protopopova, M.E. Stepanova, and A.V. Surikov. 2013. Crowdsourcing morphological annotation. http://opencorpora.org.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Cristina Bosco, Manuela Sanguinetti, and Leonardo Lesmo. 2012. The Parallel-TUT: a multilingual and multiformat treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey. ELRA.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2000. Alpino: Wide Coverage Computational Analysis of Dutch. *Computational Linguistics in the Netherlands 2000. Selected Papers from the 11th CLIN Meeting.*, pages 45–59.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2(4):597–620.

Thorsten Brants. 2000. TnT: A Statistical Part-of-speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter F Brown, Peter V DeSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18:467–479.

Matthias Buch-Kromann and Iørn Korzen. 2010. The Unified Annotation of Syntax and Discourse in the Copenhagen Dependency Treebanks. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 127–131, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie Candito, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karén Fort, Djamé Seddah, and Éric Villemonte De La Clergerie. 2014. Deep Syntax Annotation of the Sequoia French Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. ELRA.

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. *The Szeged Treebank*. Karloy Vary, Czech Republic.

Leon Derczynski, Sean Chester, and Kenneth S.Bøgh. 2015. Tune your brown clustering, please. In *Proceedings of the conference on Recent Advances in Natural Language Processing*, pages 110 – 117, Bulgaria.

Jan Einarsson. 1976. Talbankens skriftspråkskonkordans. Technical report, Lund University: Department of Scandinavian Languages.

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå Corpus (SUC) version 1.0. *Department of Linguistics, Umeå University*.

Tomaž Erjavec. 2002. Compiling and Using the IJS-ELAN Parallel Corpus. In *Informatica*, pages 299–307.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. ELRA.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. ELRA.

Eugenie Giesbrecht and Stefan Evert. 2009. Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, San Sebastian, Spain.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Sigrún Helgadóttir, Ásta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjanadóttir, and Hrafn Loftsson. 2012. The Tagged Icelandic Corpus (MIM). In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, Istanbul, Turkey. ELRA.

D. Hládek, J. Staš, and J. Juhár. 2012. Dagger: The Slovak morphological classifier. In *Proceedings ELMAR-2012*, pages 195–198.

Tobias Horsmann and Torsten Zesch. 2016. LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 120–126, Berlin, Germany.

Alon Itai and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Sara Može, Nina Ledinek, and Nanika Holz. 2013. Training corpus ssj500k 1.3. Slovenian language resource repository CLARIN.SI.

John D Lafferty, Andrew McCallum, and Fernando C N Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernández Astudillo, Silvio Amir, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Paris, France. ELRA.

Montserrat Marimon, Núria Bel, Beatriz Fisas, Blanca Arias, Silvia Vázquez, Jorge Vivaldi, Carlos Morell, and Mercè Lorente. 2014. The IULA Spanish LSP Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. ELRA.

W. Nelson Francis and Henry Kuçera. 1964. *Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers*. Department of Linguistics, Brown University.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The Dynamic Neural Network Toolkit - Technical Report.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal Dependencies 1.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague.

Robert Östling. 2013. Stagger: An Open-Source Part of Speech Tagger for Swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Olutobi Owoputi, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 NAACL: HLT*. Association for Computational Linguistics.

Patrick Paroubek. 2000. Language Resources as by-Product of Evaluation: The MULTITAG Example. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece. 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 ACL2016*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco. ELRA.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1799–1802, Genoa. ELRA.

Steffen Remus, Gerold Hintz, Darina Benikova, Thomas Arnold, Judith Eckle-Kohler, Christian M. Meyer, Margot Mieskes, and Chris Biemann. 2016. EmpiriST: AIPHES Robust Tokenization and POS-Tagging for Different Genres. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X)*, pages 106–114, Berlin, Germany.

Mojgan Seraji. 2011. A Statistical Part-of-Speech Tagger for Persian. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*.

Per Erik Solberg, Arne Skjæholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. ELRA.

Drahomíra "johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece. Association for Computational Linguistics.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Ra Kübler, and Universität Tübingen. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1175–1178. ELRA.

Atro Voutilainen. 2011. FinnTreeBank: Creating a research resource and service for language researchers with Constraint Grammar. *Constraint Grammar Applications*, page 41.