Modeling Composite Labels for Neural Morphological Tagging

Alexander Tkachenko
Institute of Computer Science
University of Tartu
Tartu, Estonia
aleksandr.tkatsenko@ut.ee

Kairit Sirts
Institute of Computer Science
University of Tartu
Tartu, Estonia
kairit.sirts@ut.ee

Abstract

Neural morphological tagging has been regarded as an extension to POS tagging task, treating each morphological tag as a monolithic label and ignoring its internal structure. We propose to view morphological tags as composite labels and explicitly model their internal structure in a neural sequence tagger. For this, we explore three different neural architectures and compare their performance with both CRF and simple neural multiclass baselines. We evaluate our models on 49 languages and show that the neural architecture that models the morphological labels as sequences of morphological category values performs significantly better than both baselines establishing state-of-the-art results in morphological tagging for most languages.[1]

1 Introduction

The common approach to morphological tagging combines the set of word's morphological features into a single monolithic tag and then, similar to POS tagging, employs multiclass sequence classification models such as CRFs (Müller et al., 2013) or recurrent neural networks (Labeau et al., 2015; Heigold et al., 2017). This approach, however, has a number of limitations. Firstly, it ignores the intrinsic compositional structure of the labels and treats two labels that differ only in the value of a single morphological category as completely independent; compare for instance labels [POS=NOUN,CASE=NOM,NUM=SG] and [POS=NOUN,CASE=NOM,NUM=PL] that only differ in the value of the NUM category. Secondly, it introduces a data sparsity issue as the less frequent labels can have only few occurrences in the training data. Thirdly, it excludes the ability to predict labels not present in the training set which can be an issue for languages such as Turkish where the number of morphological tags is theoretically unlimited (Yuret and Türe, 2006).

To address these problems we propose to treat morphological tags as composite labels and explicitly model their internal structure. We hypothesise that by doing that, we are able to alleviate the sparsity problems, especially for languages with very large tagsets such as Turkish, Czech or Finnish, and at the same time also improve the accuracy over a baseline using monolithic labels. We explore three different neural architectures to model the compositionality of morphological labels. In the first architecture, we model all morphological categories (including POS tag) as independent multiclass classifiers conditioned on the same contextual word representation. The second architecture organises these multiclass classifiers into a hierarchy—the POS tag is predicted first and the values of morphological categories are predicted conditioned on the value of the predicted POS. The third architecture models the label as a sequence of morphological category-value pairs. All our models share the same neural encoder architecture based on bidirectional LSTMs to construct contextual representations for words (Lample et al., 2016).

We evaluate all our models on 49 UD version 2.1 languages. Experimental results show that our sequential model outperforms other neural counterparts establishing state-of-the-art results in morphological tagging for most languages. We also confirm that all neural models perform significantly better than a competitive CRF baseline. In short, our contributions can be summarised as follows:

1) We propose to model the compositional internal structure of complex morphological la-

[1] The source code is available at https://github.com/AleksTk/seq-morph-tagger

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 368–379
Brussels, Belgium, October 31 - November 1, 2018.

bels for morphological tagging in a neural sequence tagging framework;
2) We explore several neural architectures for modeling the composite morphological labels;
3) We find that tag representation based on the sequence learning model achieves state-of-the art performance on many languages.
4) We present state-of-the-art morphological tagging results on 49 languages on the UDv2.1 corpora.

2 Related Work

Most previous work on modeling the internal structure of complex morphological labels has occurred in the context of morphological disambiguation—a task where the goal is to select the correct analysis from a limited set of candidates provided by a morphological analyser. The most common strategy to cope with a large number of complex labels has been to predict all morphological features of a word using several independent classifiers whose predictions are later combined using some scoring mechanism (Hajič and Hladká, 1998; Hajič, 2000; Smith et al., 2005; Yuret and Türe, 2006; Zalmout and Habash, 2017; Kirov et al., 2017). Inoue et al. (2017) combined these classifiers into a multitask neural model sharing the same encoder, and predicted both POS tag and morphological category values given the same contextual representation computed by a bidirectional LSTM. They showed that the multitask learning setting outperforms the combination of several independent classifiers on tagging Arabic. In this paper, we experiment with the same architecture, termed as multiclass multilabel model, on many languages. Additionally, we extend this approach and explore a hierarchical architecture where morphological features directly depend on the POS tag.

Another previously adopted approach involves modeling complex morphological labels as sequences of morphological feature values (Hakkani-Tur et al., 2000; Schmid and Laws, 2008). In neural networks, this idea can be implemented with recurrent sequence modeling. Indeed, one of our proposed models generates morphological tags with an LSTM network. Similar idea has been applied for the morphological reinflection task (Kann and Schütze, 2016; Faruqui et al., 2016) where the sequential model is used to generate the spellings of inflected forms given the lemma and the morphological label of the desired form. In morphological tagging, however, we generate the morphological labels themselves.

Another direction of research on modeling the structure of complex morphological labels involves structured prediction models (Müller et al., 2013; Müller and Schütze, 2015; Malaviya et al., 2018; Lee et al., 2011). Lee et al. (2011) introduced a factor graph model that jointly infers morphological features and syntactic structures. Müller et al. (2013) proposed a higher-order CRF model which handles large morphological tagsets by decomposing the full label into POS tag and morphology part. Malaviya et al. (2018) proposed a factorial CRF to model pairwise dependencies between individual features within morphological labels and also between labels over time steps for cross-lingual transfer. Recently, neural morphological taggers have been compared to the CRF-based approach (Heigold et al., 2017; Yu et al., 2017). While Heigold et al. (2017) found that their neural model with bidirectional LSTM encoder surpasses the CRF baseline, the results of Yu et al. (2017) are mixed with the convolutional encoder being slightly better or on par with the CRF but the LSTM encoder being worse than the CRF baseline.

Most previous work on neural POS and morphological tagging has shared the general idea of using bidirectional LSTM for computing contextual features for words (Ling et al., 2015; Huang et al., 2015; Labeau et al., 2015; Ma and Hovy, 2016; Heigold et al., 2017). The focus of the previous work has been mostly on modeling the inputs by exploring different character-level representations for words (Heigold et al., 2016; Santos and Zadrozny, 2014; Ma and Hovy, 2016; Inoue et al., 2017; Ling et al., 2015; Rei et al., 2016). We adopt the general encoder architecture from these works, constructing word representations from characters and using another bidirectional LSTM to encode the context vectors. In contrast to these previous works, our focus is on modeling the compositional structure of the complex morphological labels.

The morphologically annotated Universal Dependencies (UD) corpora (Nivre et al., 2017) offer a great opportunity for experimenting on many languages. Some previous work have reported results on several UD languages (Yu et al., 2017; Heigold et al., 2017). Morphological tagging results on many UD languages have been also reported for parsing systems that predict POS and morphological tags as preprocessing (Andor et al., 2016; Straka et al., 2016; Straka and Straková, 2017). Since UD treebanks have been in constant development, these results have been obtained on different UD versions and thus are not necessarily directly comparable. We conduct experiments on all UDv2.1 languages and we aim to provide a baseline for future work in neural morphological tagging.

3 Neural Models

We explore three different neural architectures for modeling morphological labels: multiclass multilabel model that predicts each category value separately, hierarchical multiclass multilabel model where the values of morphological features depend on the value of the POS, and a sequence model that generates morphological labels as sequences of feature-value pairs.

3.1 Notation

Given a sentence $w_1, \ldots, w_n$ consisting of n words, we want to predict the sequence $t_1, \ldots, t_n$ of morphological labels for that sentence. Each label $t_i = \{f_{i0}, f_{i1}, \ldots, f_{im}\}$ consists of a POS tag ($f_{i0} \equiv$ POS) and a sequence of m category values. For each word w_i, the encoder computes a contextual vector h_i, which captures information about the word and its left and right context.

3.2 Decoder Models

Multiclass Multilabel model (MCML) This model formulates the morphological tagging as a multiclass multilabel classification problem. For each morphological category, a separate multiclass classifier is trained to predict the value of that category (Figure 1 (a)). Because not all categories are always present for each POS (e.g., a noun does not have a *tense* category), we extend the morphological label of each word by adding all features that are missing from the annotated label and assign them a special value that marks the category as "off". Formally, the model can be described as:

$$p(t|h)_{\text{MCML}} = \prod_{j=0}^{M} p(f_j|h), \quad (1)$$

where M is the total number of morphological categories (such as case, number, tense, etc.) observed in the training corpus. The probability of each feature value is computed with a softmax function:

$$p(f_j|h)_{\text{MCML}} = \text{softmax}(W_j h + b_j),$$

where W_j and b_j are the parameter matrix and bias vector for the jth morphological feature ($j = 0, \ldots, M$). The final morphological label for a word is obtained by concatenating predictions for individual categories while filtering out off-valued categories.

Hierarchical Multiclass Multilabel model (HMCML) This is a hierarchical version of the MCML architecture that models the values of morphological categories as directly dependent on the POS tag (Figure 1 (b)):

$$p(t|h)_{\text{HMCML}} = p(\text{POS}|h) \prod_{j=1}^{M} p(f_j|\text{POS}, h) \quad (2)$$

The probability of the POS is computed from the context vector h using the respective parameters:

$$p(\text{POS}|h) = \text{softmax}(W_{\text{POS}} h + b_{\text{POS}})$$

The POS-dependent context vector l is obtained by concatenating the context vector h with the unnormalised log probabilities of the POS:

$$l = [h; W_{\text{POS}} h + b_{\text{POS}}]$$

The probabilities of the morphological features are computed using the POS-dependent context vector:

$$p(f_j|\text{POS}, h) = \text{softmax}(W_j l + b_j) \quad j = 1, \ldots, M$$

Sequence model (SEQ) The SEQ model predicts complex morphological labels as sequences of category values. This approach is inspired from neural sequence-to-sequence models commonly used for machine translation (Cho et al., 2014; Sutskever et al., 2014). For each word in a sentence, the decoder uses a unidirectional LSTM network (Figure 1 (c)) to generate a sequence of morphological category-value pairs based on the context vector h and the previous predictions. The probability of a morphological label t is under this model:

$$p(t|h)_{\text{SEQ}} = \prod_{j=0}^{m} p(f_j|f_0, \ldots, f_{j-1}, h) \quad (3)$$

Decoding starts by passing the start-of-sequence symbol as input. At each time step, the decoder computes the label context vector g_j based on the previously predicted category value, previous label context vector and the word's context vector.

$$g_j = \text{LSTM}([f_{j-1}; h], g_{j-1})$$

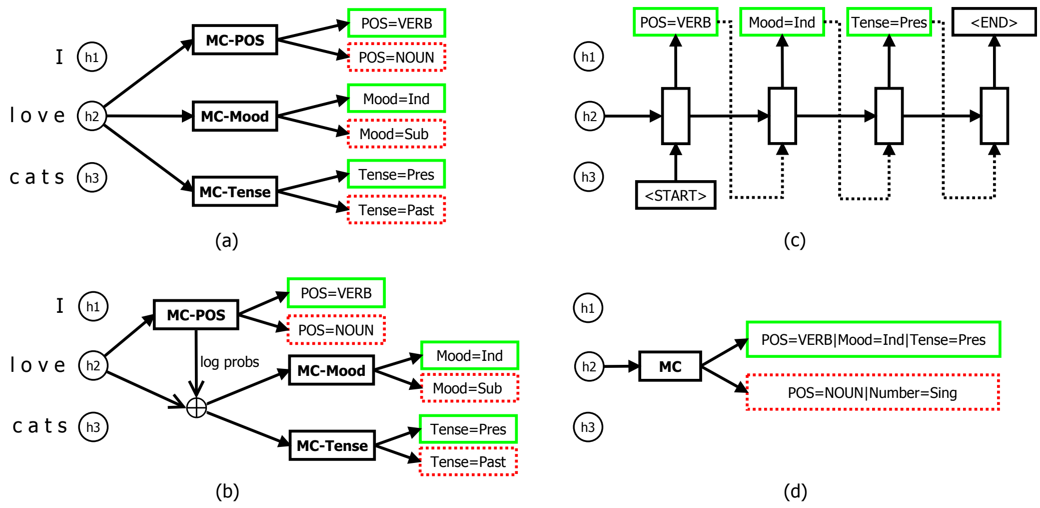

Figure 1: Neural architectures for modeling complex morphological labels: a) Multiclass Multilabel model (MCML), b) Hierarchical Multiclass Multilabel model (HMCML), c) Sequence model (SEQ) and d) Multiclass baseline model (MC). Correct labels are shown with a green border, incorrect labels have a red dotted border.

The probability of each morphological feature-value pair is then computed with a softmax.

$$p(f_j|g_j)_{\text{SEQ}} = \text{softmax}(W_{\text{SEQ}}g_j + b_{\text{SEQ}})$$

At training time, we feed correct labels as inputs while at inference time, we greedily emit the best prediction from the set of all possible feature-value pairs. The decoding terminates once the end-of-sequence symbol is produced.

3.3 Encoder

We adopt a standard sequence tagging encoder architecture for all our models. It consists of a bidirectional LSTM network that maps words in a sentence into context vectors using character and word-level embeddings. Character-level word embeddings are constructed with a bidirectional LSTM network and they capture useful information about words' morphology and shape. Word level embeddings are initialised with pre-trained embeddings and fine-tuned during training. The character and word-level embeddings are concatenated and passed as inputs to the bidirectional LSTM encoder. The resulting hidden states h_i capture contextual information for each word in a sentence. Similar encoder architectures have been applied recently with notable success to morphological tagging (Heigold et al., 2017; Yu et al., 2017) as well as several other sequence tagging tasks (Lample et al., 2016; Chiu and Nichols, 2016; Ling et al., 2015).

4 Experimental Setup

This section details the experimental setup. We describe the data, then we introduce the baseline models and finally we report the hyperparameters of the models.

4.1 Data

We run experiments on the Universal Dependencies version 2.1 (Nivre et al., 2017). We excluded corpora that did not include train/dev/test split, word form information[2], or morphological features[3]. Additionally, we excluded corpora for which pre-trained word embeddings were not available.[4] The resulting dataset contains 69 corpora covering 49 different languages. Tagsets were constructed by concatenating the POS and morphological annotations of the treebanks. Table 1 gives corpus statistics. We present type and token counts for both training and test sets. For training set, we also show the average and maximum number of tags per word type and the size of the morphological tagset. For the test set, we report the proportion of out-of-vocabulary (OOV) words as well as the number of OOV tag tokens and types.

In the encoder, we use fastText word embeddings (Bojanowski et al., 2017) pre-trained on Wikipedia.[5] Although these embeddings are uncased, our model still captures case information by

[2] French-FTB and Arabic-NYUAD
[3] Japanese
[4] Ancient Greek and Coptic
[5] https://github.com/facebookresearch/fastText

Dataset	Train set						Test set				
	Tokens	Types	Tags per word Avg	Tags per word Max	% Emb	# Tags	Tokens	Types	% OOV	OOV Tags Tokens	OOV Tags Types
Afrikaans	33894	5080	1.1	4	62.7	61	10065	2476	13.8	3	3
Arabic	254340	33225	1.8	10	90.1	349	32128	8754	9.8	6	6
Basque	72974	19222	1.4	13	53.8	884	24374	8896	17.8	71	61
Belarusian	5217	2303	1.4	6	74.6	346	1382	708	39.7	48	32
Bulgarian	124336	25047	1.1	7	65.7	432	15724	5974	12.3	4	3
Catalan	418494	31544	1.2	8	62.0	267	58017	9832	5.2	3	3
Chinese	98608	17610	1.3	6	65.8	31	12012	4055	12.5	1	1
Croatian	169283	34968	1.6	19	66.0	1105	13228	5513	14.1	13	13
Czech	1175374	125358	1.7	25	59.7	2630	174252	37727	7.0	127	94
Czech-CAC	473622	66272	1.7	21	72.4	1746	10900	4499	12.6	17	17
Czech-CLTT	27005	4336	1.5	21	73.3	418	4126	1169	17.2	39	30
Czech-FicTree	134059	25943	1.4	58	72.9	1464	16761	5691	12.8	46	43
Danish	80378	16330	1.2	5	62.3	157	10023	3424	15.3	3	2
Dutch	186046	26665	1.2	6	59.8	62	11046	3054	13.7	23	1
Dutch-LassySmall	81243	14622	1.1	5	54.7	60	10080	3573	7.4	0	0
English	204607	19672	1.4	10	58.3	117	25097	5630	9.1	3	3
English-LinES	50095	7436	1.2	4	79.8	17	15623	3530	10.3	0	0
English-ParTUT	43545	6963	1.3	8	74.7	133	3412	1136	9.3	3	3
Estonian	85567	23055	1.3	7	58.0	662	10618	4928	18.6	28	24
Finnish	162827	49210	1.1	9	59.4	2052	21070	9112	23.7	144	119
Finnish-FTB	127845	39755	1.2	8	59.3	1762	16311	8011	23.0	83	76
French	366371	42268	1.2	10	53.5	228	10298	3284	5.8	1	1
French-ParTUT	24922	3815	1.3	10	87.3	197	2693	831	11.2	2	2
French-Sequoia	51924	8463	1.2	5	73.2	200	10360	3023	8.9	0	0
Galician	86676	13236	1.1	4	73.5	27	32390	7169	9.9	3	2
Galician-TreeGal	5262	1873	1.3	9	77.7	173	10900	3182	26.8	81	41
German	268145	49472	2.3	38	25.3	684	16537	5406	11.7	28	26
Gothic	35024	6787	1.4	12	1.5	623	10182	2827	12.4	28	23
Greek	43440	9049	1.3	15	74.4	349	10922	3370	16.4	9	6
Hebrew	169360	29638	1.3	8	87.8	521	15134	5115	16.1	7	6
Hindi	281057	16974	2.4	55	79.3	939	35430	5335	4.6	23	23
Hungarian	20166	7767	1.4	5	75.7	580	10448	4558	37.1	108	85
Indonesian	97531	19223	1.2	6	45.3	21	11780	4354	13.8	0	0
Irish	3183	1257	1.5	8	62.3	236	10138	3245	36.1	276	113
Italian	288750	28915	1.2	11	70.1	278	11153	3533	5.6	0	0
Italian-ParTUT	52390	8323	1.1	6	82.0	205	3929	1318	9.1	1	1
Italian-PoSTWITA	53725	12363	1.2	9	48.7	201	6778	2550	17.3	6	4
Kazakh	547	343	1.2	2	73.2	72	10142	4559	71.9	2371	371
Korean	52328	27714	1.1	4	68.8	11	10926	7060	37.5	0	0
Latin	8018	3854	1.4	7	64.6	347	10954	4996	45.8	153	76
Latin-ITTB	270403	12526	1.5	13	63.1	985	10561	1642	2.2	14	12
Latin-PROIEL	147044	22258	1.4	21	50.6	993	12152	4331	9.8	15	13
Latvian	62397	17745	1.3	30	64.0	742	14490	5467	23.9	46	36
Lithuanian	3210	1522	1.2	3	73.2	297	1060	625	54.7	72	57
Marathi	3253	969	1.6	70	78.1	261	448	199	26.3	19	15
Norwegian-Bokmaal	243887	30072	1.2	6	61.8	203	29966	6616	11.3	4	3
Norwegian-Nynorsk	245330	29133	1.3	8	50.0	184	24773	5963	11.1	3	2
Old_Church_Slavonic	37432	7745	1.4	11	2.2	859	10031	3243	14.1	87	66
Persian	122180	13859	1.1	5	89.7	162	16122	3945	8.5	3	2
Polish	63070	21230	1.5	12	72.3	991	10906	5107	24.2	30	26
Portuguese	222070	27396	1.4	35	61.9	375	10942	3417	8.2	3	3
Portuguese-BR	273176	29944	1.2	8	58.5	22	33638	8047	6.8	0	0
Romanian	185113	30970	1.2	6	69.3	451	16324	5755	10.4	7	6
Russian	75964	25708	1.5	15	66.6	693	11548	5717	26.4	31	23
Russian-SynTagRus	871082	107891	1.4	12	74.7	723	117470	29078	9.5	14	14
Serbian	65764	14713	1.4	12	59.4	539	10891	4038	16.2	8	8
Slovak	80575	21104	1.4	39	63.7	1199	13028	6049	35.8	72	58
Slovenian	112530	29390	1.4	7	67.2	1101	14077	5856	19.9	20	19
Slovenian-SST	9487	2672	1.4	5	90.5	500	10000	2812	21.6	202	132
Spanish	389703	46979	1.4	12	56.7	399	12267	4114	7.4	3	3
Spanish-AnCora	446145	38456	1.2	8	68.3	295	52801	10615	5.6	4	2
Swedish	66645	12911	1.2	8	70.3	202	20377	5127	14.9	12	8
Swedish-LinES	48325	9659	1.2	6	77.3	168	15029	4150	15.0	875	16
Tamil	6849	3040	1.1	4	78.7	201	2183	1132	44.3	20	15
Telugu	5082	1743	1.1	4	0.3	14	721	387	25.0	0	0
Turkish	39169	14576	1.2	9	67.5	972	10256	5139	26.4	87	82
Ukrainian	75054	23970	1.4	23	72.6	1197	14939	6337	27.2	72	60
Urdu	108690	9547	2.7	52	73.9	1001	14806	2949	6.4	27	21
Vietnamese	20285	3625	1.2	4	33.7	15	11955	2684	17.1	1	1

Table 1: Descriptive statistics for all UDv2.1 datasets. For training sets we report the number of word tokens and types, the average (Avg) and maximum (Max) tags per word type, the proportion of word types for which pre-trained embeddings were available (% Emb) and the size of the morphological tagset (# Tags). For the test sets, we also give the total number of tokens and types, the proportion of OOV words (% OOV) and the number of OOV tag tokens and types.

means of character-level embeddings. In Table 1, we also report for each language the proportion of word types for which the pre-trained embeddings are available.

4.2 Baseline Models

We use two models as baseline: the CRF-based MARMOT (Müller et al., 2013) and the regular neural multiclass classifier.

MarMoT (MMT) MARMOT[6] is a CRF-based morphological tagger which has been shown to achieve competitive performance across several languages (Müller et al., 2013). MARMOT approximates the CRF objective using a pruning strategy which enables training higher-order models and handling large tagsets. In particular, the tagger first predicts the POS part of the label and based on that, constrains the set of possible morphological labels. Following the results of Müller et al. (2013), we train second-order models. We tuned the regularization type and weight on German development set and based on that, we use L2 regularization with weight 0.01 in all our experiments.

Neural Multiclass classifier (MC) As the second baseline, we employ the standard multiclass classifier used by both Heigold et al. (2017) and Yu et al. (2017). The proposed model consists of an LSTM-based encoder, identical to the one described above in section 3.3, and a softmax classifier over the full tagset. The tagset sizes for each corpora are shown in Table 1. During preliminary experiments, we also added CRF layer on top of softmax, but as this made the decoding process considerably slower without any visible improvement in accuracy, we did not adopt CRF decoding here. The multiclass model is shown in Figure 1 (d).

The inherent limitation of both baseline models is their inability to predict tags that are not present in the training corpus. Although the number of such tags in our data set is not large, it is nevertheless non-zero for most languages.

4.3 Training and Parametrisation

Since tuning model hyperparameters for each of the 69 datasets individually is computationally demanding, we optimise parameters on Finnish—a morphologically complex language with a reasonable dataset size—and apply the resulting values to other languages. We first tuned the character embedding size and character-LSTM hidden layer size of the encoder on the SEQ model and reused the obtained values with all other models. We tuned the batch size, the learning rate and the decay factor for the SEQ and MC models separately since these models are architecturally quite different. For the MCML and HMCML models we reuse the values obtained for the MC model. The remaining hyperparameter values are fixed. Table 2 lists the hyperparameters for all models.

	SEQ	OTHER NN
Encoder		
Word embedding size	300	300
Character embedding size	100	100
Character LSTM hidden layer size	150	150
Word embedding dropout	0.5	0.5
LSTM layers	1	1
LSTM hidden state size	400	400
LSTM input dropout	0.5	0.5
LSTM state dropout	0.3	0.3
LSTM output dropout	0.5	0.5
Decoder		
LSTM hidden state size	800	800
Tag embedding size	150	–
Training		
Initial learning rate	1.0	1.0
Batch size	5	20
Maximum epochs	400	400
Learning rate decay factor	–	0.98

Table 2: Hyperparameters for neural models.

We train all neural models using stochastic gradient descent for up to 400 epochs and stop early if there has been no improvement on development set within 50 epochs. For all models except SEQ, we decay the learning rate by a factor of 0.98 after every 2500 batch updates. We initialise biases with zeros and parameter matrices using Xavier uniform initialiser (Glorot and Bengio, 2010).

Words in training sets with no pre-trained embeddings are initialised with random embeddings. At test time, words with no pre-trained embedding are assigned a special UNK-embedding. We train the UNK-embedding by randomly substituting the singletons in a batch with the UNK-embedding with a probability of 0.5.

5 Results

Table 3 presents the experimental results. We report tagging accuracy for all word tokens and also for OOV tokens only. A full morphological tag is considered correct if both its POS and all morphological features are correctly predicted.

[6] http://cistern.cis.lmu.de/marmot/

	Full tag (all words)					Full tag (OOV words)					POS (all words)				
Dataset	**MMT**	**Mc**	**McMl**	**HMcMl**	**Seq**	**MMT**	**Mc**	**McMl**	**HMcMl**	**Seq**	**MMT**	**Mc**	**McMl**	**HMcMl**	**Seq**
Afrikaans	94.17	95.17	94.46	94.65	**95.45**	79.77	84.67	81.93	82.72	**84.88**	96.47	97.40	97.62	97.48	**97.66**
Arabic	90.96	93.39	93.25	93.23	**93.84**	81.25	86.06	85.24	85.14	**87.14**	95.22	96.01	96.18	96.20	**96.22**
Basque	87.15	89.92	89.96	90.15	**90.33**	63.67	**72.65**	71.61	71.86	71.95	93.87	95.25	**96.00**	96.00	95.89
Belarusian	73.66	71.35	72.29	75.54	**78.15**	48.18	46.35	47.81	52.92	**59.12**	90.38	86.54	91.53	**93.42**	93.20
Bulgarian	95.90	97.03	96.76	96.76	**97.04**	82.62	**88.74**	86.66	86.72	88.22	98.04	98.64	98.76	**98.82**	98.79
Catalan	96.60	97.52	97.39	97.36	**97.59**	89.21	91.95	91.75	91.35	**92.28**	98.05	98.63	98.68	98.65	**98.70**
Chinese	90.91	92.97	92.79	92.47	**93.27**	77.90	82.24	81.71	81.17	**82.91**	91.89	93.84	93.70	93.44	**94.11**
Croatian	84.99	88.66	88.96	88.96	**89.24**	66.11	74.87	75.89	**76.48**	76.37	96.47	97.25	**97.54**	97.41	97.45
Czech	93.00	**95.81**	95.06	95.05	95.39	73.07	**82.92**	80.81	80.53	79.70	98.56	98.95	**99.00**	98.99	98.88
Czech-CAC	90.46	**95.19**	94.74	94.72	95.14	69.39	**82.25**	80.13	79.91	81.59	98.65	99.06	99.17	**99.28**	99.05
Czech-CLTT	89.21	89.63	90.45	91.01	**91.37**	73.00	77.78	78.48	78.20	**80.03**	98.01	97.99	98.91	**99.05**	98.67
Czech-FicTree	91.24	93.93	94.54	94.48	**94.64**	75.32	83.96	84.48	83.87	**85.46**	97.55	98.14	**98.57**	98.51	98.38
Danish	93.90	95.73	95.26	95.46	**95.97**	78.74	85.24	83.03	83.68	**85.96**	95.79	97.26	97.30	97.44	**97.51**
Dutch	91.84	94.62	93.70	93.81	**94.73**	70.49	**81.23**	77.65	77.52	80.57	94.39	96.23	96.22	96.11	**96.35**
Dutch-LassySmall	97.09	97.05	97.33	97.29	**97.54**	80.73	83.96	83.15	82.35	**84.10**	97.82	97.83	**98.41**	98.36	98.26
English	93.03	**94.92**	94.40	94.36	94.80	76.22	**85.43**	83.33	83.38	84.69	94.54	**96.13**	96.09	95.96	96.06
English-LinES	95.03	**96.52**	96.36	96.39	96.36	83.72	**90.34**	89.41	90.09	89.23	95.03	**96.52**	96.36	96.39	96.36
English-ParTUT	92.32	93.76	93.17	93.17	**94.17**	70.22	76.49	73.35	73.67	**81.82**	93.87	95.43	**96.10**	96.07	95.87
Estonian	91.40	93.28	93.17	93.25	**93.30**	79.25	84.78	84.42	84.32	**85.13**	95.54	96.61	96.74	**96.85**	96.68
Finnish	91.41	93.13	93.18	93.29	**93.41**	78.35	84.05	**84.79**	84.71	84.71	95.68	96.55	97.02	**97.05**	96.79
Finnish-FTB	90.59	93.91	**94.13**	93.88	91.93	76.06	84.65	**85.50**	85.24	80.85	93.36	95.73	**96.28**	96.19	94.56
French	95.68	96.36	95.97	96.17	**96.39**	82.67	87.02	86.36	85.19	**87.85**	96.93	97.48	97.43	**97.50**	97.49
French-ParTUT	92.91	93.50	93.28	92.94	**93.95**	71.10	**73.42**	70.10	70.10	72.43	95.77	96.10	**96.77**	96.73	96.77
French-Sequoia	95.99	96.66	96.51	96.31	**96.91**	76.99	**83.64**	80.82	80.39	82.23	97.68	98.06	**98.33**	98.17	98.32
Galician	96.97	97.65	97.72	97.70	**97.76**	84.94	88.66	88.98	88.85	**89.01**	97.10	97.80	97.87	97.84	**97.90**
Galician-TreeGal	86.31	83.83	85.00	85.31	**86.61**	68.40	66.77	67.83	68.28	**71.80**	90.13	88.36	91.99	**92.00**	91.48
German	80.81	87.98	87.11	87.16	**88.32**	63.12	**78.53**	75.00	76.14	78.37	92.60	94.47	94.56	**94.62**	94.35
Gothic	87.09	86.49	86.25	86.86	**87.99**	**69.70**	65.59	60.84	62.03	65.27	95.47	94.48	95.45	**96.02**	95.59
Greek	91.00	92.63	93.85	93.58	**94.14**	73.17	78.42	80.55	79.32	**81.89**	96.74	97.21	**97.80**	97.74	97.73
Hebrew	93.19	95.05	94.73	94.60	**95.09**	81.05	87.90	86.87	86.63	**88.02**	96.15	**97.59**	97.59	97.53	97.56
Hindi	89.00	**91.78**	91.47	91.34	91.75	62.35	**72.37**	69.99	68.77	71.70	96.20	97.00	**97.32**	97.22	97.03
Hungarian	71.47	80.96	82.89	82.45	**84.12**	49.42	67.14	70.08	68.87	**72.01**	92.78	93.94	95.30	95.31	**95.44**
Indonesian	93.56	**93.79**	93.73	93.74	93.65	88.22	88.04	**88.53**	88.16	87.67	93.57	93.81	93.81	**93.85**	93.69
Irish	**67.99**	60.73	62.02	61.95	65.81	**35.48**	28.05	29.50	28.70	34.50	83.62	79.10	84.01	**84.22**	83.63
Italian	97.06	97.53	97.31	97.31	**97.61**	86.61	**88.87**	86.61	86.29	88.71	97.74	98.16	98.19	**98.32**	98.26
Italian-ParTUT	96.13	**97.12**	96.79	96.84	97.12	80.22	**90.81**	86.35	85.79	88.30	97.28	97.86	**98.14**	98.12	98.12
Italian-PoSTWITA	91.92	**93.79**	93.23	93.36	93.69	75.85	**82.34**	80.20	80.80	81.83	93.54	95.32	**95.72**	95.68	95.16
Kazakh	**37.19**	31.63	28.84	28.70	34.35	**20.97**	13.52	10.45	10.38	17.84	52.73	48.94	52.38	**54.74**	54.57
Korean	93.98	95.82	95.55	95.49	**95.87**	90.48	**93.51**	93.12	92.90	93.33	93.98	95.82	95.55	95.50	**95.87**
Latin	64.94	64.10	65.35	65.88	**67.45**	41.05	42.54	42.58	43.30	**46.99**	80.73	80.97	84.84	**85.57**	84.81
Latin-ITTB	92.98	95.18	**95.60**	95.57	95.27	68.26	74.78	**75.65**	74.35	72.61	97.30	98.12	98.30	**98.34**	98.17
Latin-PROIEL	88.37	**90.64**	90.20	90.13	89.66	68.43	**78.39**	74.46	73.20	71.69	95.78	96.68	**96.80**	96.72	95.94
Latvian	85.59	87.67	87.14	87.14	**87.79**	67.91	73.59	71.94	71.94	**73.88**	92.80	94.38	94.87	**94.88**	94.55
Lithuanian	65.00	58.02	64.91	63.58	**67.92**	44.66	36.72	43.79	43.10	**51.03**	73.87	70.00	81.60	79.25	**81.70**
Marathi	66.07	68.75	64.06	64.96	**70.09**	39.83	**49.15**	33.05	36.44	44.92	82.14	82.81	84.15	**84.82**	84.60
Norwegian-Bokmaal	94.99	96.37	96.13	95.94	**96.54**	80.14	84.53	83.11	82.54	**84.68**	97.33	98.24	98.39	98.26	**98.44**
Norwegian-Nynorsk	94.65	**96.25**	95.69	95.69	96.07	81.32	**85.30**	81.93	82.11	83.82	97.08	98.12	**98.22**	98.14	98.08
Old_Church_Slavonic	87.58	86.96	87.01	86.87	**87.96**	60.31	**60.59**	57.49	57.13	58.83	94.98	94.40	95.38	**95.61**	94.94
Persian	95.84	96.75	96.38	96.38	**96.79**	79.36	**86.09**	84.04	83.67	85.43	96.39	97.13	97.11	97.10	**97.30**
Polish	86.04	90.46	**90.99**	90.78	90.99	69.13	**81.21**	79.36	79.81	80.87	96.65	97.73	**98.25**	98.11	98.04
Portuguese	94.21	95.59	95.34	95.59	**95.75**	79.48	86.66	86.66	**86.77**	86.43	97.21	97.72	**98.06**	97.95	98.04
Portuguese-BR	97.59	98.20	98.20	98.14	**98.21**	92.30	95.20	**95.56**	95.03	95.16	97.60	98.20	98.21	98.16	**98.22**
Romanian	96.30	97.00	96.72	96.61	**97.16**	85.15	89.51	88.10	87.92	**89.75**	97.18	97.61	97.74	**97.78**	97.77
Russian	85.99	90.21	90.73	90.93	**91.05**	66.91	77.85	78.24	78.90	**79.26**	95.42	96.43	96.72	**96.84**	96.50
Russian-SynTagRus	94.44	**96.78**	96.48	96.58	96.67	78.91	**88.50**	87.21	87.48	86.98	98.51	98.84	98.92	98.93	**98.94**
Serbian	91.17	93.25	93.32	93.58	**93.93**	77.32	83.22	82.20	82.48	**83.50**	97.47	97.89	**98.25**	98.17	98.19
Slovak	81.72	87.50	88.16	**88.54**	88.46	68.42	78.66	78.98	79.24	**79.69**	94.62	95.85	**96.49**	96.34	96.46
Slovenian	89.39	94.32	94.05	93.98	**94.62**	73.14	86.34	83.94	83.58	**86.41**	97.07	98.15	98.29	98.39	**98.42**
Slovenian-SST	78.71	75.75	79.18	80.02	**80.44**	45.45	44.06	48.40	49.88	**52.24**	88.44	87.54	92.04	**92.38**	90.99
Spanish	94.33	**95.05**	94.82	94.81	94.90	77.34	**82.95**	82.29	82.18	81.52	95.88	96.89	96.95	**96.98**	96.83
Spanish-AnCora	97.13	**97.67**	97.54	97.58	97.63	90.26	93.22	93.09	93.19	**93.36**	98.25	98.64	98.75	**98.78**	98.68
Swedish	94.28	95.41	95.07	95.25	**95.65**	82.72	86.11	84.20	84.07	**86.37**	96.38	97.49	97.69	**97.72**	97.66
Swedish-LinES	85.24	86.38	85.99	85.98	**86.47**	64.01	**68.33**	66.28	65.79	67.26	95.00	96.17	**96.69**	96.65	96.25
Tamil	81.40	82.18	83.05	81.26	**85.75**	67.87	71.90	72.42	70.56	**75.83**	86.39	87.49	**91.07**	90.24	90.75
Telugu	**92.23**	90.43	89.04	89.32	91.26	**80.00**	75.56	70.00	71.67	78.33	**92.23**	90.43	89.04	89.32	91.26
Turkish	86.09	89.47	90.69	90.51	**90.70**	63.97	74.85	**79.83**	79.02	79.13	92.86	94.67	**95.54**	95.51	95.19
Ukrainian	85.33	88.98	89.94	**89.96**	89.81	69.19	78.89	79.24	79.34	**79.36**	95.97	96.40	**97.23**	97.06	97.03
Urdu	77.37	80.09	79.52	78.54	**80.66**	54.99	64.54	60.30	61.68	**65.07**	92.56	93.29	**93.87**	93.71	93.81
Vietnamese	86.13	**88.66**	88.51	88.22	88.44	55.19	**70.81**	70.46	69.29	68.70	86.15	**88.67**	88.58	88.34	88.46
Average (>100K)	92.18	**94.37**	94.12	94.07	**94.37**	76.65	**84.03**	82.67	82.47	83.42	96.49	97.37	**97.52**	97.50	97.40
Average (50K-100K)	91.27	93.36	93.36	93.40	**93.66**	76.96	**83.46**	82.39	82.43	83.40	95.39	96.43	**96.71**	96.67	96.65
Average (20K-50K)	87.56	89.42	89.69	89.66	**90.43**	66.35	72.55	71.76	71.47	**73.84**	94.37	95.13	95.83	**95.86**	95.69
Average (<20K)	71.35	68.68	69.37	69.65	**72.78**	49.19	47.46	46.58	47.52	**53.26**	82.07	80.22	84.27	84.60	**84.70**
Overall average	88.18	89.58	89.61	89.64	**90.42**	71.12	76.74	75.62	75.64	**77.52**	93.76	94.27	95.11	**95.14**	95.08

Table 3: Morphological tagging accuracies on UDv2.1 test sets for MarMot (MMT) and MC baselines as well as for MCML, HMCML and SEQ compositional models. The left section shows the full POS+MORPH tag results, the middle section gives accuracies for OOV words only, the right-most section shows the POS tagging accuracy. The best result in each section for each language is in bold. The languages are color-coded according to the training set size, lighter color denotes larger training set: cyan (<20K), violet (20K-50K), magenta (50K-100K), pink (>100K).

Feature	Seq	Mc	#	Feature	Seq	Mc	#
POS	**91.03**	90.20	69	NumType	**89.68**	87.82	54
Number	**94.02**	93.05	63	Polarity	**93.83**	92.86	54
VerbForm	**91.29**	89.86	61	Degree	**87.44**	84.12	48
Person	**89.02**	87.52	60	Poss	**94.52**	93.60	44
Tense	**92.96**	91.31	59	Voice	**88.40**	82.85	42
PronType	**89.83**	88.81	58	Definite	**95.26**	94.10	37
Mood	**87.34**	85.40	58	Aspect	**89.76**	87.71	29
Gender	**89.31**	87.78	55	Animacy	**86.22**	83.73	19
Case	**88.90**	87.04	55	Polite	75.76	**80.48**	10

Table 4: Performance of Seq and Mc models on individual features reported as macro-averaged F1-scores.

First of all, we can confirm the results of Heigold et al. (2017) that the performance of neural morphological tagging indeed exceeds the results of a CRF-based model. In fact, all our neural models perform significantly better than MarMoT ($p < 0.001$).[7]

The best neural model on average is the Seq model, which is significantly better from both the Mc baseline as well as the other two compositional models, whereby the improvement is especially well-pronounced on smaller datasets. We do not observe any significant differences between McMl and HMcMl models neither on all words nor OOV evaluation setting.

We also present POS tagging results in the rightmost section of Table 3. Here again, all neural models are better than CRF which is in line with the results presented by Plank et al. (2016). For POS tags, the HMcMl is the best on average. It is also significantly better than the neural Mc baseline, however, the differences with the McMl and Seq models are insignificant.

In addition to full-tag accuracies, we assess the performance on individual features. Table 4 reports macro-averaged F1-cores for the Seq and the Mc models on universal features. Results indicate that the Seq model systematically outperforms the Mc model on most features.

6 Analysis and Discussion

OOV label accuracy Our models are able to predict labels that were not seen in the training data. Figure 2 presents the accuracy of test tokens with OOV labels obtained with our best performing Seq model plotted against the number of OOV label types. The datasets with zero accuracy are omitted. The main observation is that although the OOV label accuracy is zero for some languages, it is above zero on ca. half of the datasets—a result that would be impossible with MarMoT or Mc baselines.

[7]As indicated by Wilcoxon signed-rank test.

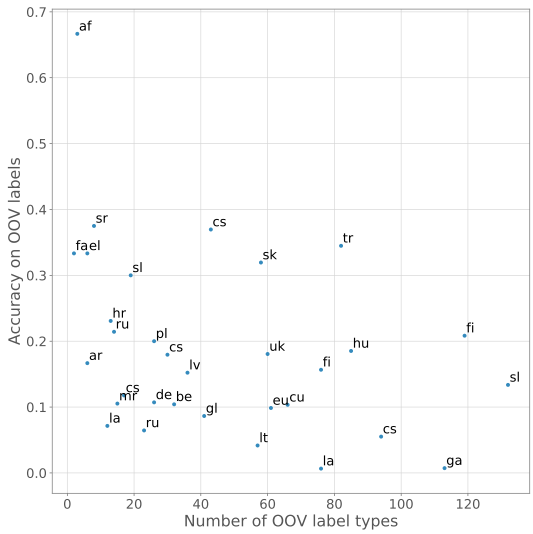

Figure 2: OOV label accuracies of the Seq model.

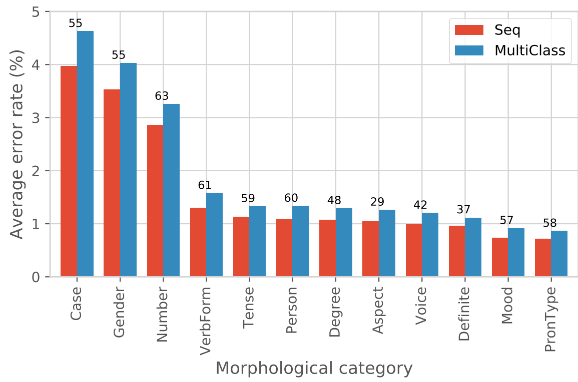

Figure 3: Average error rates of distinct morphological categories for Seq and Mc models.

Error Analysis Figure 3 shows the largest error rates for distinct morphological categories for both Seq and Mc models averaged over all languages. We observe that the error patterns are similar for both models but the error rates of the Seq model are consistently lower as expected.

Stability Analysis To assess the stability of our predictions, we picked five languages from different families and with different corpus size, and performed five independent train/test runs for each language. Table 5 summarises the results of these experiments and demonstrates a reasonably small variance for all languages. For all languages, except for Finnish, the worst accuracy of the Seq model was better than the best accuracy of the Mc model, confirming our results that in those languages, the Seq model is consistently better than the Mc baseline.

Dataset	SEQ	MC
Finnish	93.24 ± 0.12	93.20 ± 0.07
German	88.45 ± 0.21	87.74 ± 0.17
Hungarian	84.51 ± 0.54	80.68 ± 0.48
Russian	91.08 ± 0.18	90.13 ± 0.15
Turkish	90.29 ± 0.24	89.16 ± 0.27

Table 5: Mean accuracy with standard deviation over five independent runs for SEQ and MC models.

Hyperparameter Tuning It is possible that the hyperparameters tuned on Finnish are not optimal for other languages and thus, tuning hyperparameters for each language individually would lead to different conclusions than currently drawn. To shed some light on this issue, we tuned hyperparameters for the SEQ and MC models on the same subset of five languages. We first independently optimised the dropout rates on word embeddings, encoder's LSTM inputs and outputs, as well as the number of LSTM layers. We then performed a grid search to find the optimal initial learning rate, the learning rate decay factor and the decay step. Value ranges for the tuned parameters are given in Table 6.

Parameter	Values
Word embedding dropout	$\{0, 0.1, \ldots, 0.5\}$
LSTM input dropout	$\{0, 0.1, \ldots, 0.5\}$
LSTM input dropout	$\{0, 0.1, \ldots, 0.5\}$
Number of LSTM layers	$\{1, 2\}$
Initial learning rate	$\{0.01, 0.1, 1, 2\}$
Learning rate decay factor	$\{0.97, 0.98, 0.99, 1\}$
Decay step	$\{1250, 2500, 5000\}$

Table 6: The grid values for hyperparameter tuning.

Table 7 reports accuracies for the tuned models compared to the mean accuracies reported in Table 5. As expected, both tuned models demonstrate superior performance on all languages, except for German with the SEQ model. Hyperparameter tuning has a greater overall effect on the MC model, which suggests that it is more sensitive to the choice of parameters than the SEQ model. Still, the tuned SEQ model performs better or at least as good as the MC model on all languages.

Comparison with Previous Work Since UD datasets have been in rapid development and different UD versions do not match, direct comparison of our results to previously published results is difficult. Still, we show the results taken from Heigold et al. (2017), which were obtained on UDv1.3, to provide a very rough comparison. In addition, we compare our SEQ model with a neural tagger presented by Dozat et al. (2017), which is similar to our MC model, but employs a more sophisticated encoder. We train this model on UDv2.1 on the same set of languages used by Heigold et al. (2017).

Table 8 reports evaluation results for the three models. The SEQ model and Dozat's tagger demonstrate comparable performance. This suggests that the SEQ model can be further improved by adopting a more advanced encoder from Dozat et al. (2017).

Dataset	SEQ	Gain	MC	Gain
Finnish	93.44	+0.20	93.43	+0.23
German	88.35	−0.10	88.14	+0.40
Hungarian	85.56	+1.05	82.29	+1.61
Russian	91.44	+0.36	90.74	+0.61
Turkish	90.56	+0.27	89.32	+0.16

Table 7: Accuracies of the tuned SEQ and MC models compared to the mean accuracies in Table 5.

Dataset	SEQ	Dozat	Heigold
Arabic	**93.84**	92.85	93.78
Bulgarian	97.04	**97.25**	95.14
Czech	**95.39**	95.22	96.32
English	94.80	**94.81**	93.32
Estonian	93.30	**93.90**	94.25
Finnish	93.41	**93.73**	93.52
French	**96.39**	95.90	94.91
Hindi	91.75	**92.36**	90.84
Hungarian	**84.12**	82.84	77.59
Romanian	97.16	**97.20**	94.12
Russian-SynTagRus	**96.67**	96.20	96.45
Turkish	**90.70**	90.22	89.12
Average	**93.71**	93.54	92.45

Table 8: Accuracies for the SEQ model, Dozat et al. (2017) and Heigold et al. (2017).

7 Conclusion

We hypothesised that explicitly modeling the internal structure of complex labels for morphological tagging improves the overall tagging accuracy over the baseline with monolithic tags. To test this hypothesis, we experimented with three approaches to model composite morphological tags in a neural sequence tagging framework. Experimental results on 49 languages demonstrated the advantage of modeling morphological labels as sequences of category values, whereas the superiority of this model is especially pronounced on smaller datasets. Furthermore, we showed that, in contrast to baselines, our models are capable of predicting labels that were not seen during training.

Acknowledgments

This work was supported by the Estonian Research Council (grants no. 2056, 1226 and IUT34-4).

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 2442–2452. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5:135–146.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association of Computational Linguistics*, 4:357–370.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Proceedings of the conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task." Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256.

Jan Hajič. 2000. Morphological tagging: Data vs. dictionaries. In *Proceedings of the 1st Conference of the North American Chapter of the Association of Computational Linguistics*, pages 94–101. Association for Computational Linguistics.

Jan Hajič and Barbora Hladká. 1998. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 1, pages 483–490. Association for Computational Linguistics.

Diiek Z Hakkani-Tur, Kemal Oflazer, and Gokhan Tur. 2000. Statistical morphological disambiguation for agglutinative languages. In *Proceedings of the 18th International Conference on Computational Linguistics*, volume 1.

Georg Heigold, Josef van Genabith, and Günter Neumann. 2016. Scaling character-based morphological tagging to fourteen languages. In *2016 IEEE International Conference on Big Data*, pages 3895–3902. IEEE.

Georg Heigold, Guenter Neumann, and Josef van Genabith. 2017. An extensive empirical evaluation of character-based morphological tagging for 14 languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 505–513.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Go Inoue, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Joint prediction of morphosyntactic categories for fine-grained arabic part-of-speech tagging exploiting tag dictionary information. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 421–431.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 555–560. Association for Computational Linguistics.

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for english: Exploring the cross-linguistic tradeoff between morphology and syntax. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 112–117.

Matthieu Labeau, Kevin Löser, and Alexandre Allauzen. 2015. Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 232–237.

Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

John Lee, Jason Naradowsky, and David A Smith. 2011. A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association* *for Computational Linguistics: Human Language Technologies-Volume 1*, pages 885–894. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1064–1074. Association for Computational Linguistics.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663. Association for Computational Linguistics.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order crfs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, PhÆřÆąng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, LÆřÆąng Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Robert Östling, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Benoît Sagot, Shadi Saleh, Tanja Samardžić, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Jonathan North Washington, Mats Wirén, Tak-sum Wong, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 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.

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*, page 412. Association for Computational Linguistics.

Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1818–1826.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 777–784. Association for Computational Linguistics.

Noah A Smith, David A Smith, and Roy W Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 475–482. Association for Computational Linguistics.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Xiang Yu, Agnieszka Falenska, and Ngoc Thang Vu. 2017. A general-purpose tagger with convolutional neural networks. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 124–129.

Deniz Yuret and Ferhan Türe. 2006. Learning morphological disambiguation rules for turkish. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 328–334. Association for Computational Linguistics.

Nasser Zalmout and Nizar Habash. 2017. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for arabic. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 704–713.