Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing

Milan Straka and **Jana Straková** and **Jan Hajič**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{strakova,straka,hajic}@ufal.mff.cuni.cz

Abstract

We present an extensive evaluation of three recently proposed methods for contextualized embeddings on 89 corpora in 54 languages of the Universal Dependencies 2.3 in three tasks: POS tagging, lemmatization, and dependency parsing. Employing the BERT, Flair and ELMo as pretrained embedding inputs in a strong baseline of UDPipe 2.0, one of the best-performing systems of the CoNLL 2018 Shared Task and an overall winner of the EPE 2018, we present a one-to-one comparison of the three contextualized word embedding methods, as well as a comparison with word2vec-like pretrained embeddings and with end-to-end character-level word embeddings. We report state-of-the-art results in all three tasks as compared to results on UD 2.2 in the CoNLL 2018 Shared Task.

1 Introduction

We publish a comparison and evaluation of three recently proposed contextualized word embedding methods: BERT (Devlin et al., 2018), Flair (Akbik et al., 2018) and ELMo (Peters et al., 2018), in 89 corpora which have a training set in 54 languages of the Universal Dependencies 2.3 in three tasks: POS tagging, lemmatization and dependency parsing. Our contributions are the following:

- Meaningful massive comparative evaluation of BERT (Devlin et al., 2018), Flair (Akbik et al., 2018) and ELMo (Peters et al., 2018) contextualized word embeddings, by adding them as input features to a strong baseline of UDPipe 2.0, one of the best performing systems in the CoNLL 2018 Shared Task (Zeman et al., 2018) and an overall winner of the EPE 2018 Shared Task (Fares et al., 2018).
- State-of-the-art results in POS tagging, lemmatization and dependency parsing in UD 2.2, the dataset used in CoNLL 2018 Shared Task (Zeman et al., 2018).
- We report our best results on UD 2.3. The addition of contextualized embeddings improvements range from 25% relative error reduction for English treebanks, through 20% relative error reduction for high resource languages, to 10% relative error reduction for all UD 2.3 languages which have a training set.

2 Related Work

A new type of deep contextualized word representation was introduced by Peters et al. (2018). The proposed embeddings, called ELMo, were obtained from internal states of deep bidirectional language model, pretrained on a large text corpus. Akbik et al. (2018) introduced analogous contextual string embeddings called Flair, which were obtained from internal states of a *character-level* bidirectional language model. The idea of ELMos was extended by Devlin et al. (2018), who instead of a bidirectional recurrent language model employ a Transformer (Vaswani et al., 2017) architecture.

The *Universal Dependencies*[1] project (Nivre et al., 2016) seeks to develop cross-linguistically consistent treebank annotation of morphology and syntax for many languages. The latest version UD 2.3 (Nivre et al., 2018) consists of 129 treebanks in 76 languages, with 89 of the treebanks containing a train a set and being freely available. The annotation consists of UPOS (universal POS tags), XPOS (language-specific POS tags), Feats (universal morphological features), Lemmas, dependency heads and universal dependency labels.

In 2017 and 2018, CoNLL Shared Tasks *Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2017, 2018) were held in order to stimulate research in multi-lingual POS

[1] https://universaldependencies.org/

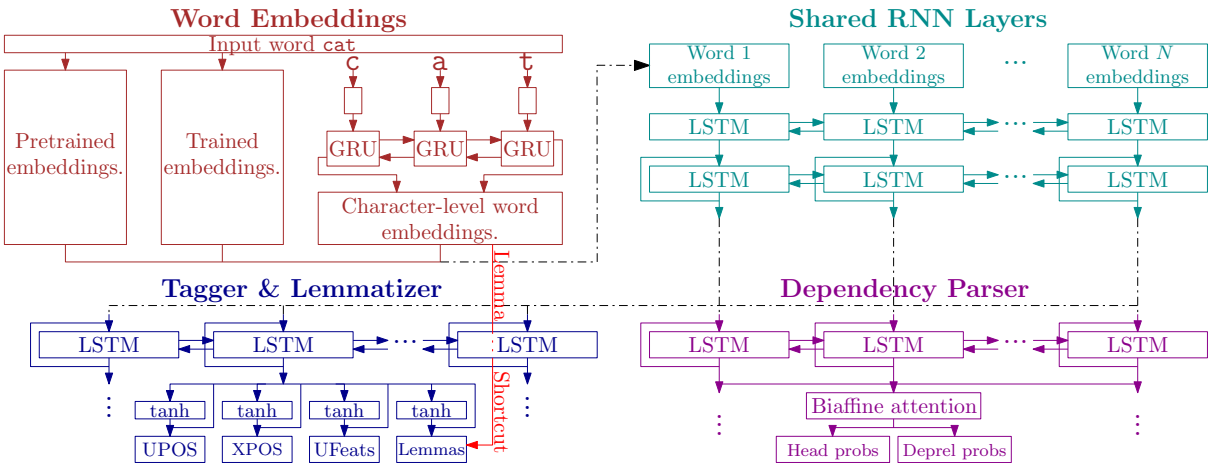


Figure 1: UDPipe 2.0 architecture overview.

tagging, lemmatization and dependency parsing.

The system of Che et al. (2018) is one of the three winners of the CoNLL 2018 Shared Task. The authors employed manually trained ELMo-like contextual word embeddings, reporting 7.9% error reduction in LAS parsing performance.

3 Methods

Our **baseline** is the *UDPipe 2.0* (Straka, 2018) participant system from the CoNLL 2018 Shared Task (Zeman et al., 2018). The system is available at `http://github.com/CoNLL-UD-2018/UDPipe-Future`.

A graphical overview of the UDPipe 2.0 is shown in Figure 1. In short, UDPipe 2.0 is a multi-task model predicting POS tags, lemmas and dependency trees jointly. After embedding input words, two shared bidirectional LSTM (Hochreiter and Schmidhuber, 1997) layers are performed. Then, tagger and lemmatizer specific bidirectional LSTM layer is executed, with softmax classifiers processing its output and generating UPOS, XPOS, Feats and Lemmas. The lemmas are generated by classifying into a set of edit scripts which process input word form and produce lemmas by performing character-level edits on the word prefix and suffix. The lemma classifier additionally takes the character-level word embeddings as input.

Finally, the output of the two shared LSTM layers is processed by a parser specific bidirectional LSTM layer, whose output is then passed to a biaffine attention layer (Dozat and Manning, 2016) producing labeled dependency trees. We refer the readers for detailed treatment of the architecture and the training procedure to Straka (2018).

The simplest baseline system uses only end-to-end word embeddings trained specifically for the task. Additionally, the UDPipe 2.0 system also employs the following two embeddings:

- **word embeddings (WE):** We use FastText word embeddings (Bojanowski et al., 2017) of dimension 300, which we pretrain for each language on Wikipedia using segmentation and tokenization trained from the UD data.[2]
- **character-level word embeddings (CLE):** We employ bidirectional GRUs of dimension 256 in line with Ling et al. (2015): we represent every Unicode character with a vector of dimension 256, and concatenate GRU output for forward and reversed word characters. The character-level word embeddings are trained together with UDPipe network.

Optionally, we add pretrained contextual word embeddings as another input to the neural network. Contrary to finetuning approach used by the BERT authors (Devlin et al., 2018), we never finetune the embeddings.

- **BERT** (Devlin et al., 2018): We employ three pretrained models of dimension 768:[3] an English one for the English treebanks (`Base Uncased`), a Chinese one for Chinese and Japanese treebanks (`Base Chinese`) and a multilingual one (`Base Multilingual Uncased`) for all other languages. We produce embedding of a UD word as an average of BERT subword embeddings this UD word

[2] We use `-minCount 5 -epoch 10 -neg 10` options and keep at most one million most frequent words.

[3] From `https://github.com/google-research/bert`.

WE	CLE	Bert	UPOS	XPOS	UFeats	Lemma	UAS	LAS	MLAS	BLEX
			90.14	88.51	86.50	88.64	79.43	73.55	56.52	60.84
WE			94.91	93.51	91.89	92.10	85.98	81.73	68.47	70.64
	CLE		95.75	94.69	93.43	96.24	86.99	82.96	71.06	75.78
WE	CLE		96.39	95.53	94.28	96.51	87.79	84.09	73.30	77.36
		Base	96.35	95.08	93.56	93.29	89.31	85.69	74.11	75.45
WE		Base	96.62	95.54	94.08	93.77	89.49	85.96	74.94	76.27
	CLE	Base	96.86	95.96	94.85	96.64	89.76	86.29	76.20	79.87
WE	CLE	Base	**97.00**	**96.17**	**94.97**	**96.66**	**89.81**	**86.42**	**76.54**	**80.04**

Table 1: BERT Base compared to word embeddings (WE) and character-level word embeddings (CLE). Results for 72 UD 2.3 treebanks with train and development sets and non-empty Wikipedia.

Language	Bert	UPOS	XPOS	UFeats	Lemma	UAS	LAS	MLAS	BLEX
English	Base	**97.38**	**96.97**	97.22	**97.71**	**91.09**	**88.22**	**80.48**	**82.38**
English	Multi	97.36	**96.97**	**97.29**	97.63	90.94	88.12	80.43	82.22
Chinese	Base	**97.07**	**96.89**	**99.58**	99.98	**90.13**	**86.74**	**79.67**	**83.85**
Chinese	Multi	96.27	96.25	99.37	**99.99**	87.58	83.96	76.26	81.04
Japanese	Base	**98.24**	**97.89**	99.98	**99.53**	**95.55**	**94.27**	**87.64**	**89.24**
Japanese	Multi	98.17	97.71	**99.99**	99.51	95.30	93.99	87.17	88.77

Table 2: Comparison of multilingual and language-specific BERT models on 4 English treebanks (each experiment repeated 3 times), and on Chinese-GSD and Japanese-GSD treebanks.

was decomposed into, and we average the last four layers of the BERT model.

- **Flair** (Akbik et al., 2018): Pretrained contextual word embeddings of dimension 4096 for available languages.[4]
- **ELMo** (Peters et al., 2018): Pretrained contextual word embeddings of dimension 512, available only for English.

We evaluate the metrics defined in Zeman et al. (2018) using the official evaluation script.[5] When reporting results for multiple treebanks, we compute macro-average of their scores (following the CoNLL 2018 Shared Task).

4 Results

Table 1 displays results for 72 UD 2.3 treebanks with train and development sets and non-empty Wikipedia (raw corpus for the WE), considering WE, CLE and Base BERT embeddings. Both WE and CLE bring substantial performance boost, with CLE providing larger improvements, especially for lemmatization and morphological features. Combining WE and CLE shows that the improvements are complementary and using both embeddings yields further increase.

Employing only the BERT embeddings results in significant improvements, compared to both WE and CLE individually, with highest increase for syntactic parsing, less for morphology and worse performance for lemmatization than CLE. Considering BERT versus WE+CLE, BERT offers higher parsing performance, comparable UPOS accuracy, worse morphological features and substantially lower lemmatization performance. We therefore conclude that the representation computed by BERT captures higher-level syntactic and possibly even semantic meaning, while providing less information about morphology and orthographical composition required for lemmatization.

Combining BERT and CLE results in an increased performance, especially for morphological features and lemmatization. The addition of WE provides minor improvements in all metrics, suggesting that the BERT embeddings encompass substantial amount of information which WE adds to CLE. In total, adding BERT embeddings to a baseline with WE and CLE provides a 16.9% relative error reduction for UPOS tags, 12% for mor-

[4] Models available in Jan 2018, for languages `bg`, `cs`, `de`, `en`, `fr`, `nl`, `pl`, `pt`, `sl`, `sv`.

[5] http://universaldependencies.org/conll18/conll18_ud_eval.py

WE	CLE	Bert	Flair	UPOS	XPOS	UFeats	Lemmas	UAS	LAS	MLAS	BLEX
				92.77	89.59	88.88	91.52	82.59	77.89	61.52	65.89
WE				96.63	94.48	94.01	94.82	88.55	85.25	73.38	75.74
	CLE			96.80	95.11	94.64	97.31	88.88	85.51	74.37	78.87
WE	CLE			97.32	95.88	95.44	97.62	89.55	86.46	76.42	80.36
		Base		97.49	95.68	95.17	95.45	91.48	88.69	78.61	80.14
WE		Base		97.65	96.11	95.58	95.86	91.59	88.84	79.30	80.79
	CLE	Base		97.79	96.45	95.94	97.75	91.74	88.98	79.97	83.43
WE	CLE	Base		97.89	96.58	96.09	97.78	91.80	89.09	80.30	83.59
			Flair	97.69	96.22	95.69	96.49	90.43	87.57	77.91	80.06
WE			Flair	97.77	96.37	95.87	96.62	90.53	87.69	78.37	80.37
	CLE		Flair	97.72	96.40	95.94	97.77	90.58	87.74	78.47	81.94
WE	CLE		Flair	97.76	96.50	96.06	97.85	90.66	87.83	78.73	82.16
WE	CLE	Base	Flair	**98.00**	**96.80**	**96.30**	**97.87**	**91.92**	**89.32**	**80.78**	**83.96**

Table 3: Flair compared to word embeddings (WE), character-level word embeddings (CLE) and BERT Base.

phological features, 4.3% for lemmatization, and 14.5% for labeled dependency parsing.

The influence of multilingual and language-specific BERT models is analyzed in Table 2. Surprisingly, averaged results of the four English treebanks show very little decrease when using the multilingual BERT model compared to English-specific one, most likely owing to the fact that English is the largest language used to train the multilingual model. Contrary to English, the Chinese BERT model shows substantial improvements compared to a multilingual model when utilized on the Chinese-GSD treebank, and minor improvements on the Japanese-GSD treebank.

Note that according to the above comparison, the substantial improvements offered by BERT embeddings can be achieved using a *single multilingual model*, opening possibilities for interesting language-agnostic approaches.

4.1 Flair

Table 3 shows the experiments in which WE, CLE, Flair and BERT embeddings are added to the baseline, averaging results for 23 UD 2.3 treebanks for which the Flair embeddings were available.

Comparing Flair and BERT embeddings, the former demonstrates higher performance in POS tagging, morphological features, and lemmatization, while achieving worse results in dependency parsing, suggesting that Flair embeddings capture more morphological and orthographical information. A comparison of Flair+WE+CLE with BERT+WE+CLE shows that the introduction of WE+CLE embeddings to BERT encompasses nearly all information of Flair embeddings, as demonstrated by BERT+WE+CLE achieving better performance in all tasks but lemmatization, where it is only slightly behind Flair+WE+CLE.

The combination of all embeddings produces best results in all metrics. In total, addition of BERT and Flair embeddings to a baseline with WE and CLE provides a 25.4% relative error reduction for UPOS tags, 18.8% for morphological features, 10% for lemmatization and 21% for labeled dependency parsing.

4.2 ELMo

Given that pretrained ELMo embeddings are available for English only, we present results for ELMo, Flair, and BERT contextualized embeddings on four macro-averaged English UD 2.3 treebanks in Table 4.

Flair and BERT results are consistent with the previous experiments. Employing solely ELMo embeddings achieves best POS tagging and lemmatization compared to using only BERT or Flair, with dependency parsing performance higher than Flair, but lower than BERT. Therefore, ELMo embeddings seem to encompass the most morphological and ortographical features compared to BERT and Flair, more syntactical features than Flair, but less than BERT.

When comparing ELMo with Flair+WE+CLE, the former surpass the latter in all metrics but lemmatization (and lemmatization performance is equated when employing ELMo+WE+CLE).

WE	CLE	Bert	Flair	Elmo	UPOS	XPOS	UFeats	Lemmas	UAS	LAS	MLAS	BLEX
					92.31	91.18	92.11	93.67	82.16	77.27	63.00	66.20
WE					95.69	95.30	96.15	96.27	86.98	83.59	73.29	75.40
	CLE				95.50	95.04	95.65	97.06	86.86	83.10	72.60	75.53
WE	CLE				96.33	95.86	96.44	97.32	87.83	84.52	75.08	77.65
		Base			96.88	96.46	96.94	96.18	90.98	87.98	79.66	79.94
WE		Base			97.04	96.66	97.07	96.38	91.19	88.20	80.08	80.41
	CLE	Base			97.21	96.82	97.08	97.61	91.23	88.32	80.42	82.38
WE	CLE	Base			97.38	96.97	97.22	97.70	91.09	88.22	80.48	82.38
			Flair		96.88	96.45	96.99	97.01	89.50	86.42	78.03	79.36
WE			Flair		97.06	96.56	97.03	97.12	89.68	86.67	78.55	79.85
	CLE		Flair		97.00	96.52	97.04	97.57	89.75	86.72	78.56	80.56
WE	CLE		Flair		97.02	96.55	97.12	97.63	89.67	86.64	78.41	80.48
				Elmo	97.23	96.83	97.25	97.13	90.15	87.26	79.47	80.49
WE				Elmo	97.24	96.84	97.28	97.12	90.25	87.34	79.49	80.57
	CLE			Elmo	97.21	96.81	97.23	97.62	90.22	87.30	79.51	81.32
WE	CLE			Elmo	97.21	96.82	97.27	97.63	90.33	87.42	79.66	81.50
WE	CLE	Base	Flair		**97.45**	**97.08**	97.36	**97.76**	**91.25**	**88.45**	**80.94**	**82.79**
WE	CLE	Base		Elmo	97.42	97.05	97.41	97.68	91.09	88.26	80.81	82.48
WE	CLE	Base	Flair	Elmo	97.44	**97.08**	**97.43**	97.67	91.08	88.28	80.76	82.47

Table 4: ELMo, Flair and BERT contextualized word embeddings for four macro-averaged English UD 2.3 treebanks. All experiments were performed three times and averaged.

System	UPOS	XPOS	UFeats	Lemmas	UAS	LAS	MLAS	BLEX
UDPipe 2.0 WE+CLE	95.84	94.96	94.24	95.89	85.53	82.11	72.12	75.74
UDPipe 2.0 WE+CLE+BERT	96.23	95.43	94.74	96.03	87.33	84.20	75.15	78.30
UDPipe 2.0 WE+CLE+BERT 3-model ensemble	**96.32**	**95.55**	**94.90**	**96.16**	**87.64**	**84.60**	**75.76**	**78.88**
Original UDPipe 2.0 ST entry (Straka, 2018)	*95.73*	*94.79*	*94.11*	*95.12*	*85.28*	*81.83*	*71.71*	*74.67*
HIT-SCIR Harbin (Che et al., 2018) 3-model ensemble	*96.23*	*95.16*	*91.20*	*93.42*	*87.61*	*84.37*	*70.12*	*75.05*
HIT-SCIR Harbin (Che et al., 2018) w/o ensembling						*83.75*		
Stanford (Qi et al., 2018)	*95.93*	*94.95*	*94.14*	*95.25*	*86.56*	*83.03*	*72.67*	*75.46*
TurkuNLP (Kanerva et al., 2018)	*95.41*	*94.47*	*93.82*	*96.08*	*85.32*	*81.85*	*71.27*	*75.83*

Table 5: CoNLL 2018 UD Shared Task results on treebanks with development sets (so called *big treebanks* in the shared task).

Furthermore, morphological feature generation performance of ELMo is better than BERT+WE+CLE. These results indicate that ELMo capture a lot of information present in WE+CLE, which is further promoted by the fact that ELMo+WE+CLE shows very little improvements compared to ELMo only (with the exception of lemmatization profiting from CLE).

Overall, the best-performing model on English treebanks is BERT+Flair+WE+CLE, with the exception of morphological features, where ELMo helps marginally. The relative error reduction compared to WE+CLE range from 30.5% for UPOS tagging, 26% for morphological features, 16.5% for lemmatization and 25.4% for labeled dependency parsing.

4.3 CoNLL 2018 Shared Task Results

Given that the inputs in the CoNLL 2018 Shared Task are raw texts, we reuse tokenization and segmentation employed by original UDPipe 2.0. Also, we pretrain WE not only on Wikipedia, but on all plaintexts provided by the shared tasks organizers. The resulting F1 scores of UDPipe 2.0 WE+CLE and WE+CLE+BERT on treebanks with development sets (so called *big treebanks* in the shared task) are presented in Table 5.

The inclusion of BERT embeddings results in state-of-the-art single-model performance in UPOS, XPOS, UFeats, MLAS, and BLEX met-

rics, and state-of-the-art ensemble performance in all metrics.

4.4 BERT and Flair Improvement Levels

To investigate which languages benefit most from BERT embeddings, Figure 2 presents relative error reductions in UPOS tagging, lemmatization, and unlabeled and labeled dependency parsing, as a function of logarithmic size of the respective Wikipedia (which corresponds to the size of BERT Multilingual model training data). The results indicate that consistently with intuition, larger amount of data used to pretrain the BERT model leads to higher performance.

To compare BERT and Flair embeddings, Figure 3 displays relative error improvements of Flair+WE+CLE, BERT+WE+CLE and BERT+Flair+WE+CLE models compared to WE+CLE, this time as a function of logarithmic training data size. Generally the relative error reduction decrease with the increasing amount of training data. Furthermore, the difference between Flair and BERT is clearly visible, with BERT excelling in dependency parsing and Flair in lemmatization.

4.5 UD 2.3 Detailed Performance

Table 6 shows a detailed evaluation of all 89 freely available UD 2.3 treebanks with a train set, comparing the WE+CLE baseline to the best performing WE+CLE+BERT+Flair (where Flair available) model.

The evaluation includes also 13 treebanks whose languages are not part of BERT Multilingual model. For these treebanks, the effect of using BERT embeddings is mixed, as can be observed in the Table 6 indicating which UD languages were not part of BERT training. UPOS tagging, unlabeled and labeled dependency parsing profits from BERT embedding utilization, with averaged relative error reduction of 3.8%, 2%, and 0.8%, respectively. On the other hand, lemmatization performance deteriorates, with -2.2% averaged relative error reduction.

Averaged across all treebanks, relative error improvement of BERT+Flair embeddings inclusion is 15% for UPOS tagging, 2.4% for lemmatization and 11.5% for labeled dependency parsing.

5 Conclusions

We presented a thorough evaluation of the BERT, Flair, and ELMo contextualized embeddings in 89 languages of the UD in POS tagging, lemmatization, and dependency parsing. We conclude that addition of any of the contextualized embeddings as additional inputs to a neural network results in substantial performance increase. Our findings show that the BERT embeddings yield the greatest improvements, reaching state-of-the-art results in CoNLL 2018 Shared Task and contain most complementary information as compared to word- and character-level word embeddings, while Flair embeddings encompass the morphological and orthographical information.

Acknowledgements

The work described herein has been supported by OP VVV VI LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project CZ.02.1.01/0.0/0.0/16_013/0001781) and it has been supported and has been using language resources developed by the LINDAT/CLARIN project of the the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

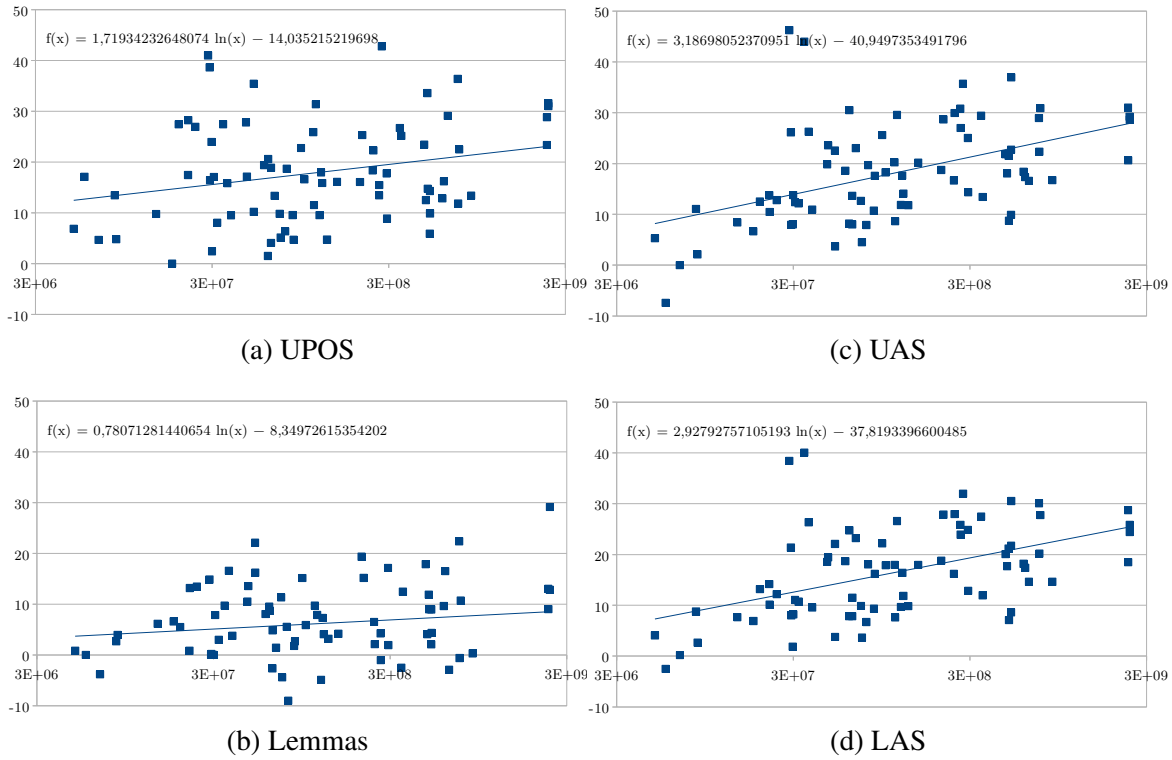
References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

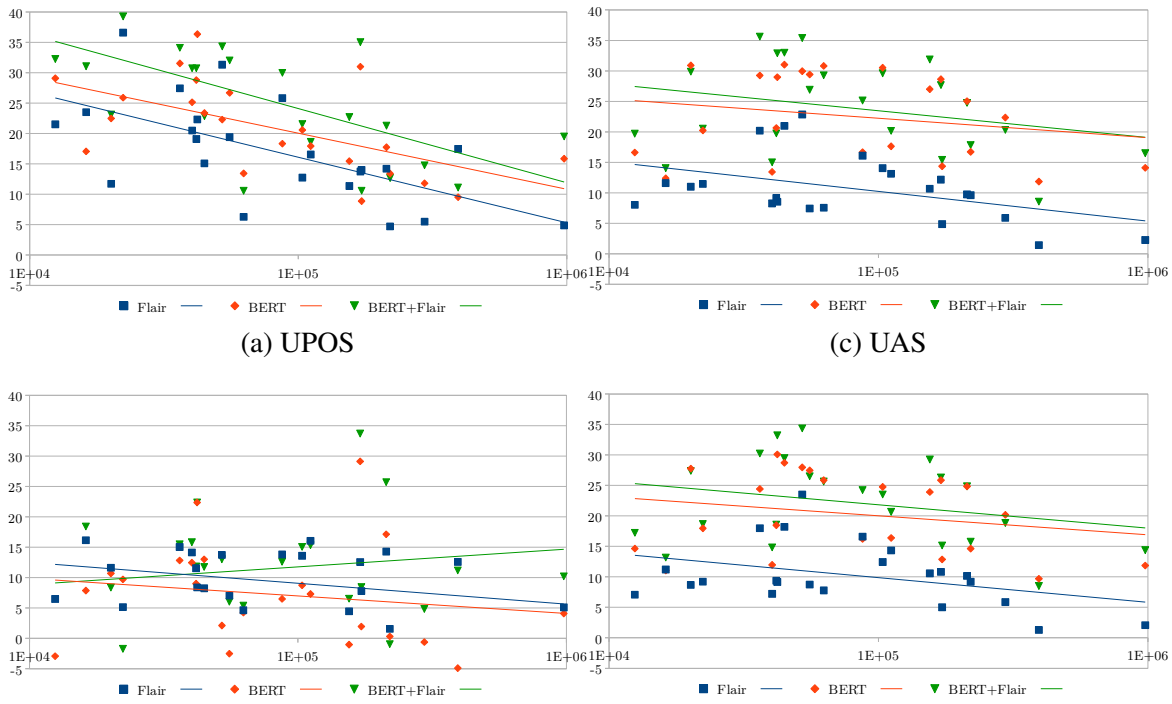
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.



(a) UPOS (c) UAS

(b) Lemmas (d) LAS

Figure 2: Relative error improvements on UD 2.3 treebanks which have a training set and their language is included in BERT model. The baseline model uses WE and CLE, and the improved model also uses BERT Multilingual contextualized embeddings. The value on the x-axis is the logarithmic size of the corresponding Wikipedia, which corresponds to training data size of the BERT Multilingual model.



(a) UPOS (c) UAS

(b) Lemmas (d) LAS

Figure 3: Relative error improvements of the baseline with WE+CLE and a model additionally including Flair and/or BERT Multilingual contextual embeddings. The value on the x-axis is the logarithmic UD train data size.

Language	BERT Train	UDPipe 2.0 with WE+CLE								UDPipe 2.0 with WE+CLE+BERT+Flair where available							
		UPOS	XPOS	UFeats	Lemmas	UAS	LAS	MLAS	BLEX	UPOS	XPOS	UFeats	Lemmas	UAS	LAS	MLAS	BLEX
Afrikaans-AfriBooms		98.25	94.48	97.66	97.46	89.38	86.58	77.66	77.82	**98.73**	**95.82**	**98.49**	**97.60**	**90.71**	**88.35**	**81.14**	**80.17**
Ancient Greek-PROIEL	✗	**97.86**	**98.08**	**92.44**	**93.51**	**85.93**	**82.11**	**67.16**	**71.22**	97.75	97.99	92.29	93.26	85.87	82.08	66.89	70.68
Ancient Greek-Perseus	✗	**93.27**	**86.22**	**91.39**	**85.02**	**78.85**	**73.54**	**53.87**	**53.19**	92.95	85.46	90.94	84.59	78.55	72.96	52.92	52.62
Arabic-PADT		96.83	93.97	94.11	95.28	87.54	82.94	73.92	75.87	**96.98**	**94.57**	**94.72**	**95.43**	**89.01**	**84.62**	**76.28**	**77.81**
Armenian-ArmTDP		93.49	—	82.85	92.86	78.62	71.27	48.11	60.11	**95.30**	—	**86.89**	**93.61**	**82.86**	**76.60**	**56.15**	**65.53**
Basque-BDT		96.11	—	92.48	96.29	86.11	82.86	72.33	78.54	**96.48**	—	**93.32**	**96.43**	**87.63**	**84.50**	**74.91**	**80.10**
Belarusian-HSE		93.63	89.80	73.30	87.34	78.58	72.72	46.20	58.28	**96.24**	**93.27**	**79.67**	**89.22**	**88.49**	**83.21**	**58.44**	**69.11**
Bulgarian-BTB		98.98	97.00	97.82	97.94	93.38	90.35	83.63	84.42	**99.20**	**97.57**	**98.22**	**98.25**	**95.34**	**92.62**	**87.00**	**87.59**
Buryat-BDT	✗	40.34	—	32.40	**58.17**	32.60	**18.83**	1.26	**6.49**	**45.50**	—	**33.49**	57.42	**35.88**	18.28	**1.48**	5.82
Catalan-AnCora		98.88	98.88	98.37	99.07	93.22	91.06	84.48	86.18	**99.06**	**99.06**	**98.60**	**99.25**	**94.49**	**92.74**	**87.36**	**88.90**
Chinese-GSD		94.88	94.72	99.22	**99.99**	84.64	80.50	71.04	76.78	**97.07**	**96.89**	**99.58**	99.98	**90.13**	**86.74**	**79.67**	**83.85**
Coptic-Scriptorium	✗	**94.72**	**93.52**	96.27	95.53	**85.69**	**81.08**	64.65	68.65	94.55	93.15	**96.44**	**95.73**	85.10	80.52	**65.16**	**68.81**
Croatian-SET		98.13	—	92.25	97.27	91.10	86.78	73.61	81.19	**98.45**	—	**93.27**	**97.64**	**93.20**	**89.35**	**77.08**	**84.44**
Czech-CAC		99.37	96.66	96.34	98.57	92.99	90.71	84.30	87.18	**99.44**	**96.94**	**96.62**	**98.73**	**93.59**	**91.50**	**85.84**	**88.47**
Czech-CLTT		98.88	91.18	91.59	**98.25**	86.90	84.03	71.63	79.20	**99.32**	**92.67**	**92.88**	98.22	**89.59**	**87.01**	**75.53**	**82.13**
Czech-FicTree		98.55	95.04	95.87	98.63	92.91	89.75	81.04	85.49	**98.82**	**96.16**	**96.88**	**98.84**	**94.34**	**91.87**	**84.80**	**88.16**
Czech-PDT		99.18	97.28	97.23	99.02	93.33	91.31	86.15	88.60	**99.34**	**97.71**	**97.67**	**99.12**	**94.43**	**92.56**	**88.09**	**90.22**
Danish-DDT		97.78	—	97.33	97.52	86.88	84.31	76.29	78.51	**98.21**	—	**97.77**	**97.72**	**89.32**	**87.24**	**80.58**	**81.93**
Dutch-Alpino		96.83	94.80	96.33	97.09	91.37	88.38	77.28	79.82	**97.55**	**95.87**	**97.34**	**97.28**	**94.12**	**91.78**	**83.12**	**84.42**
Dutch-LassySmall		96.50	95.08	96.42	97.41	90.20	86.39	77.19	78.83	**96.87**	**95.91**	**96.97**	**97.55**	**93.07**	**89.88**	**82.00**	**83.26**
English-EWT		96.29	96.10	97.10	98.25	89.63	86.97	79.00	82.36	**97.59**	**97.41**	**97.82**	**98.84**	**92.50**	**90.40**	**84.41**	**87.03**
English-GUM		96.02	95.90	96.82	96.85	87.27	84.12	73.51	74.68	**96.93**	**96.73**	**97.59**	**97.22**	**91.47**	**88.80**	**80.14**	**80.62**
English-LinES		96.91	95.62	96.31	96.45	84.15	79.71	71.38	73.22	**97.86**	**96.94**	**97.48**	**96.87**	**87.28**	**83.48**	**77.45**	**78.36**
English-ParTUT		96.10	95.83	95.51	97.74	90.29	87.27	76.44	80.33	**97.43**	**97.25**	**96.54**	**98.09**	**93.75**	**91.12**	**81.74**	**85.13**
Estonian-EDT		97.64	98.27	96.23	95.30	88.00	85.18	78.72	78.51	**97.83**	**98.36**	**96.42**	**95.44**	**89.46**	**86.77**	**80.62**	**80.17**
Finnish-FTB		96.65	95.39	96.62	95.49	90.68	87.89	80.58	81.18	**96.97**	**95.61**	**96.73**	**95.57**	**91.68**	**89.02**	**82.25**	**82.69**
Finnish-TDT		97.45	98.12	95.43	91.45	89.88	87.46	80.43	76.64	**97.57**	**98.24**	**95.80**	**91.68**	**91.66**	**89.49**	**82.89**	**78.57**
French-GSD		97.63	—	97.13	98.35	90.65	88.06	79.76	82.39	**97.98**	—	**97.42**	**98.43**	**92.55**	**90.31**	**82.66**	**85.09**
French-ParTUT		96.93	96.47	94.43	95.70	92.17	89.63	75.22	78.07	**97.64**	**97.35**	**95.12**	**96.06**	**94.51**	**92.47**	**80.50**	**82.19**
French-Sequoia		98.79	—	98.09	98.57	92.37	90.73	84.51	85.93	**99.32**	—	**98.62**	**98.89**	**94.88**	**93.81**	**89.10**	**90.08**
French-Spoken		95.91	97.30	—	**96.92**	82.90	77.53	68.24	69.47	**97.23**	**97.48**	—	96.75	**86.27**	**81.40**	**73.26**	**73.36**
Galician-CTG		97.84	97.47	**99.83**	98.58	86.44	83.82	72.46	77.21	**98.06**	**97.70**	**99.83**	**98.81**	**86.94**	**84.43**	**73.72**	**78.33**
Galician-TreeGal		95.82	92.46	93.96	97.06	82.72	77.69	63.73	68.89	**97.30**	**95.01**	**96.03**	**97.71**	**86.62**	**82.62**	**72.29**	**76.24**
German-GSD		94.48	97.31	90.68	**96.80**	85.53	81.07	58.82	72.13	**95.18**	**97.95**	**91.72**	96.77	**88.11**	**84.06**	**63.33**	**75.44**
Gothic-PROIEL	✗	96.66	**97.23**	**90.77**	**94.72**	85.27	79.60	66.71	**72.86**	**96.72**	97.22	90.58	94.47	**85.53**	**79.69**	**66.86**	72.52
Greek-GDT		97.98	97.99	94.96	95.82	92.10	89.79	78.60	79.72	**98.25**	**98.25**	**95.76**	**95.88**	**93.92**	**92.16**	**82.29**	**82.14**
Hebrew-HTB		97.02	97.03	95.87	97.12	89.70	86.86	75.52	78.14	**97.50**	**97.50**	**96.18**	**97.24**	**91.78**	**89.22**	**78.85**	**80.80**
Hindi-HDTB		97.52	97.04	94.15	**98.67**	94.85	91.83	78.49	86.83	**97.58**	**97.19**	**94.24**	**98.67**	**95.56**	**92.50**	**79.32**	**87.66**
Hungarian-Szeged		95.76	—	91.75	95.05	84.04	79.73	67.63	73.63	**97.09**	—	**93.41**	**95.44**	**88.76**	**85.12**	**74.08**	**79.21**
Indonesian-GSD		93.69	94.19	95.58	99.64	85.31	78.99	67.74	76.38	**94.09**	**94.93**	**96.03**	**99.66**	**86.47**	**80.40**	**70.01**	**78.19**
Irish-IDT		92.72	91.44	82.43	90.48	80.39	72.34	46.49	55.32	**93.22**	**92.00**	**83.78**	**90.56**	**81.43**	**73.47**	**49.05**	**56.50**
Italian-ISDT		98.39	98.30	98.11	98.66	93.49	91.54	84.28	85.49	**98.62**	**98.54**	**98.26**	**98.78**	**94.97**	**93.38**	**87.14**	**88.10**
Italian-ParTUT		98.38	98.35	97.77	98.16	92.64	90.47	81.87	82.99	**98.54**	**98.52**	**98.05**	**98.24**	**95.36**	**93.38**	**86.57**	**87.30**
Italian-PoSTWITA		96.61	96.43	96.90	97.00	86.03	81.78	72.88	74.33	**97.11**	**96.98**	**97.12**	**97.27**	**87.25**	**83.07**	**74.70**	**76.27**
Japanese-GSD		98.13	97.81	**99.98**	99.52	95.06	93.73	86.37	88.04	**98.24**	**97.89**	**99.98**	**99.53**	**95.55**	**94.27**	**87.64**	**89.24**
Kazakh-KTB		55.84	52.06	40.40	63.96	53.30	33.38	4.82	15.10	**63.08**	**60.63**	**43.64**	**64.03**	**57.02**	**38.72**	**7.88**	**18.78**
Korean-GSD		96.29	90.39	99.77	93.40	87.70	84.24	79.74	76.35	**96.99**	**91.21**	**99.83**	**93.72**	**89.38**	**86.05**	**82.19**	**78.58**
Korean-Kaist		95.59	87.00	—	**94.30**	88.42	86.48	80.72	79.22	**95.77**	**87.46**	—	94.15	**89.35**	**87.54**	**82.12**	**80.18**
Kurmanji-MG	✗	53.38	51.42	41.53	**69.58**	**45.22**	**34.32**	2.74	**19.39**	**58.78**	**56.11**	**42.03**	68.21	43.74	32.99	**3.10**	17.98
Latin-ITTB		98.34	96.37	96.97	98.99	91.06	88.80	82.35	85.71	**98.42**	**96.45**	**97.05**	**99.03**	**91.25**	**89.10**	**82.80**	**86.05**
Latin-PROIEL		97.01	97.15	91.53	**96.32**	**83.34**	78.66	**67.40**	**73.65**	**97.15**	**97.21**	**91.54**	96.18	**83.34**	**78.70**	67.29	73.52
Latin-Perseus		88.40	74.58	79.10	81.45	71.20	61.28	41.58	45.09	**89.96**	**76.22**	**80.43**	**81.95**	**74.39**	**64.68**	**44.96**	**47.94**
Latvian-LVTB		**96.11**	88.69	93.01	95.46	87.20	83.35	71.92	76.64	**96.11**	**89.06**	**93.30**	**95.76**	**88.05**	**84.50**	**73.81**	**78.33**
Lithuanian-HSE		81.70	79.91	60.47	**76.89**	51.98	42.17	18.17	28.70	**88.77**	**86.04**	**66.70**	**76.89**	**64.53**	**54.53**	**26.35**	**34.76**
Maltese-MUDT	✗	95.99	95.69	—	—	84.65	79.71	66.75	71.49	**96.15**	**95.85**	—	—	**85.31**	**80.10**	**67.21**	**71.62**
Marathi-UFAL		80.10	—	67.23	**81.31**	**70.63**	**61.41**	29.34	**45.87**	**83.50**	—	**67.96**	**81.31**	68.45	60.44	**29.58**	43.75
North Sami-Giella	✗	92.61	93.78	**90.00**	**88.34**	78.39	73.60	62.29	61.45	**92.76**	**94.11**	89.83	88.25	**78.47**	**73.95**	**62.47**	**61.68**
Norwegian-Bokmaal		98.31	—	97.14	98.64	92.39	90.49	84.06	86.53	**98.59**	—	**97.54**	**98.72**	**93.78**	**92.19**	**86.72**	**88.60**
Norwegian-Nynorsk		93.87	—	91.57	96.06	80.09	75.04	63.72	68.22	**95.52**	—	**93.17**	**96.59**	**82.64**	**78.08**	**67.53**	**71.75**
Norwegian-NynorskLIA		89.59	—	86.13	93.93	68.08	60.07	44.47	50.98	**92.53**	—	**88.96**	**94.73**	**71.42**	**64.12**	**49.10**	**55.36**
Old Church Slavonic-PROIEL	✗	96.89	**97.16**	**90.72**	**93.07**	89.64	84.99	73.66	77.71	**96.96**	97.13	90.45	92.91	**89.88**	**85.21**	**73.77**	**77.88**
Old French-SRCMF	✗	96.09	96.00	97.82	—	91.75	**86.82**	**79.89**	**83.81**	**96.26**	**96.21**	**97.89**	—	**91.83**	86.75	79.79	83.55
Persian-Seraji		97.75	97.70	97.78	**97.44**	90.05	86.66	81.23	80.93	**98.17**	**98.05**	**98.13**	97.21	**92.01**	**89.07**	**84.36**	**83.40**
Polish-LFG		98.80	94.56	95.49	97.54	96.58	94.76	87.04	90.26	**99.16**	**95.91**	**96.57**	**97.85**	**97.44**	**96.03**	**90.14**	**92.09**
Polish-SZ		98.34	93.25	93.04	97.16	93.39	91.24	81.06	85.99	**98.91**	**95.12**	**95.08**	**97.53**	**95.73**	**94.25**	**86.66**	**89.89**
Portuguese-Bosque		97.07	—	96.40	98.46	91.36	89.04	76.67	83.06	**97.38**	—	**96.96**	**98.59**	**92.69**	**90.70**	**79.59**	**85.44**
Portuguese-GSD		98.31	98.30	99.92	99.30	93.01	91.63	85.96	86.94	**98.67**	**98.67**	**99.93**	**99.48**	**94.74**	**93.71**	**89.19**	**90.28**
Romanian-Nonstandard		96.68	92.11	90.88	**94.78**	89.12	84.20	65.93	73.44	**96.85**	**92.27**	**91.04**	94.55	**89.61**	**84.78**	**66.82**	**73.77**
Romanian-RRT		97.96	97.43	97.53	98.41	91.31	86.74	79.02	81.09	**98.16**	**97.56**	**97.75**	**98.59**	**92.41**	**88.05**	**81.04**	**82.89**
Russian-GSD		97.10	96.98	92.66	97.37	88.15	84.37	74.07	80.03	**97.78**	**97.64**	**94.76**	**97.84**	**90.74**	**87.51**	**79.13**	**83.97**
Russian-SynTagRus		99.12	—	97.57	98.53	93.80	92.32	87.91	89.17	**99.23**	—	**97.97**	**98.59**	**94.92**	**93.68**	**89.85**	**90.81**
Russian-Taiga		93.18	**99.98**	82.87	89.99	75.45	69.11	48.81	57.21	**95.47**	**99.98**	**86.87**	**91.18**	**80.74**	**75.65**	**57.16**	**63.65**
Serbian-SET		98.33	—	94.35	97.36	92.70	89.27	79.14	84.18	**98.71**	—	**95.79**	**97.76**	**94.57**	**91.65**	**83.03**	**87.24**
Slovak-SNK		96.83	86.14	90.82	96.40	89.82	86.90	74.00	81.37	**97.70**	**88.54**	**93.07**	**96.75**	**94.30**	**92.15**	**81.43**	**87.24**
Slovenian-SSJ		98.61	95.70	95.92	98.25	92.96	91.16	83.85	86.89	**98.83**	**96.53**	**96.77**	**98.54**	**94.81**	**93.49**	**87.58**	**90.04**
Slovenian-SST		93.79	86.12	86.28	95.17	73.51	67.51	52.67	60.32	**95.72**	**89.25**	**89.43**	**96.06**	**77.23**	**71.79**	**58.69**	**64.84**
Spanish-AnCora		98.91	98.92	98.49	99.17	92.34	90.26	83.97	85.51	**99.05**	**99.06**	**98.70**	**99.25**	**93.75**	**92.03**	**87.03**	**88.35**
Spanish-GSD		96.85	—	97.09	98.97	90.71	88.03	75.98	81.47	**97.36**	—	**97.19**	**99.14**	**92.32**	**90.11**	**79.29**	**84.92**
Swedish Sign Language-SSLC	✗	68.44	57.27	—	—	49.82	37.94	31.34	39.47	**72.34**	**70.92**	—	—	**56.03**	**42.02**	**34.50**	**43.19**
Swedish-LinES		96.78	94.75	89.43	97.03	86.07	81.86	66.48	77.38	**97.77**	**95.97**	**90.39**	**97.50**	**88.16**	**84.55**	**70.13**	**80.81**
Swedish-Talbanken		97.94	96.71	96.86	98.01	89.63	86.61	79.67	82.26	**98.60**	**97.62**	**97.69**	**98.13**	**92.42**	**90.16**	**84.56**	**86.19**
Tamil-TTB		91.05	83.81	87.28	93.92	74.11	66.37	55.31	59.58	**92.61**	**86.53**	**89.89**	**93.97**	**77.68**	**71.14**	**60.67**	**64.74**
Telugu-MTG		93.07	93.07	**99.03**	—	91.26	85.02	77.75	**81.76**	**94.73**	**94.73**	**99.03**	—	**91.96**	**85.30**	**77.79**	81.60
Turkish-IMST		96.01	95.12	92.55	96.01	74.19	67.56	56.96	61.37	**96.07**	**95.37**	**93.25**	**96.39**	**76.30**	**70.11**	**59.91**	**64.07**
Ukrainian-IU		97.59	92.66	92.66	97.23	88.29	85.25	73.81	79.10	**98.20**	**94.63**	**94.43**	**97.65**	**91.65**	**89.36**	**79.97**	**84.24**
Upper Sorbian-UFAL	✗	62.93	—	41.10	**68.68**	45.58	34.54	3.37	16.65	**69.69**	—	**43.46**	66.80	**48.64**	**38.85**	**5.03**	**17.80**
Urdu-UDTB		93.66	91.98	81.92	97.40	87.50	81.62	55.02	73.07	**94.28**	**92.37**	**82.47**	**97.56**	**88.55**	**83.03**	**56.58**	**75.05**
Uyghur-UDT	✗	**89.87**	**92.54**	**88.30**	**95.31**	78.46	67.09	47.84	57.08	89.58	92.27	88.29	95.30	**79.10**	**67.46**	**48.09**	**57.69**
Vietnamese-VTB		89.68	87.41	**99.72**	99.55	70.38	62.56	55.56	59.54	**90.87**	**88.87**	99.68	**99.79**	**72.94**	**65.41**	**58.97**	**62.64**
Total		93.71	92.52	90.56	94.35	84.23	79.59	67.36	72.05	**94.71**	**93.69**	**91.50**	**94.51**	**86.34**	**82.01**	**70.66**	**74.75**

Table 6: Results on all UD 2.3 treebanks with a train set, comparing inclusion of BERT and possibly Flair embeddings to WE+CLE baseline. Gold tokenization and segmentation is used.

Timothy Dozat and Christopher D. Manning. 2016. Deep Biaffine Attention for Neural Dependency Parsing. *CoRR*, abs/1611.01734.

Murhaf Fares, Stephan Oepen, Lilja Øvrelid, Jari Björne, and Richard Johansson. 2018. The 2018 Shared Task on Extrinsic Parser Evaluation: On the Downstream Utility of English Universal Dependency Parsers. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 22–33. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Brussels, Belgium. Association for Computational Linguistics.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Joakim Nivre et al. 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pages 197–207, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, and Milan Straka. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.