End-to-end mBERT based Seq2seq Enhanced Dependency Parser with Linguistic Typology knowledge

Chinmay Choudhary
National University of Ireland, Galway
c.choudhary1@nuigalway.ie

Colm O'riordan
National University of Ireland, Galway
colm.oriordan@nuigalway.ie

Abstract

We describe the NUIG solution for *IWPT 2021 Shared Task* of Enhanced Dependency (ED) parsing in multiple languages. For this shared task, we propose and evaluate an End-to-end Seq2seq mBERT- based ED parser which predicts the ED-parse tree of a given input sentence as a relative head-position tag-sequence. Our proposed model is a multitasking neural-network which performs five key tasks simultaneously namely *UPOS-tagging*, *UFeat-tagging*, *Lemmatization*, *Dependency-parsing* and *ED-parsing*. Furthermore we utilise the linguistic typology available in the **WALS** database to improve the ability of our proposed end-to-end parser to transfer across languages. Results show that our proposed *Seq2seq ED-parser* performs on par with state-of-the-art *ED-parser* despite having a much simpler design.

1 Introduction

The Enhanced Universal Dependency (EUD) Parsing (Schuster and Manning, 2016; Nivre et al., 2020) framework is an interesting extension of the standard Dependency Parsing framework, which provides additional significant syntactic and semantic knowledge, that is missing in a standard dependency parse-tree. Such additional knowledge can be crucial for numerous downstream NLP tasks.

The *IWPT 2021 Shared Task* (Bouma et al., 2021) requires the participants to perform the enhanced dependency parsing of the given test-sentences, in addition to predicting the sentence-boundaries, token-boundaries, lemmatization, POS-tags, morphological features and the basic dependency relations. The participants are provided with the blind test-corpora in 17 languages, and are expected to perform the enhanced dependency parsing on each sentence within these test corpora and submit the results (in the conllu format).

For this *IWPT 2021 Shared Task* (Bouma et al., 2021) we propose and evaluate the performance an *End-to-end mBERT Based Se2seq ED-Parser* which performs five key tasks namely *UPOS-tagging*, *UFeats-prediction*, *Lemmatization*, *Dependency-parsing* and *Enhanced Dependency-parsing* in multi-tasking settings.

Our proposed model is an extension of the popular UDify model (Kondratyuk and Straka, 2019) which is the state-of-the-art mBERT based multilingual dependency parser, and is inspired by (Li et al., 2018) which is an End-to-end Seq2seq Dependency-Parser. We describe the UDify model in Section 2.

We trained our proposed ED-Parser on a large joint polyglot corpus created by concatenating all the treebanks in the provided training dataset for *IWPT 2021 Shared Task*, and evaluated it on eight of the 17 provided blind test-corpora.

Furthermore, similar to previous approaches (Ammar et al., 2016), we utilized the *Linguistic Typology* knowledge available in **World Atlas of Language System (WALS)** database (Haspelmath, 2009) to improve the cross-lingual transferring ability of our proposed ED-parser. We fed these typology features together with token-ids into the proposed ED-parser. We describe the architecture of our *End-to-end mBERT Based Se2seq ED-Parser* in detail in Section 3.

2 Background and Related Work

2.1 Seq2seq Dependency Parser

(Li et al., 2018) proposed a Seq2seq architecture to perform the end-to-end dependency parsing. The approach represented the entire dependency parse-tree of a given input-sentence, as a relative head-position tag-seq (of same length as the length of the input sentence). Figure 1 depicts a labelled and an unlabelled parse-tree represented by their re-

Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021), pages 225–232
Bangkok, Thailand (online), August 6, 2021.

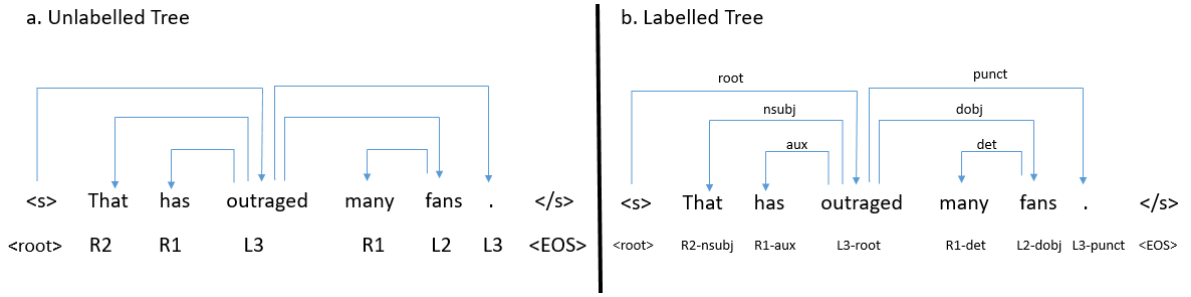

Figure 1: Examples of dependency parse tree being represented as relative head-position tag sequence by (Li et al., 2018)

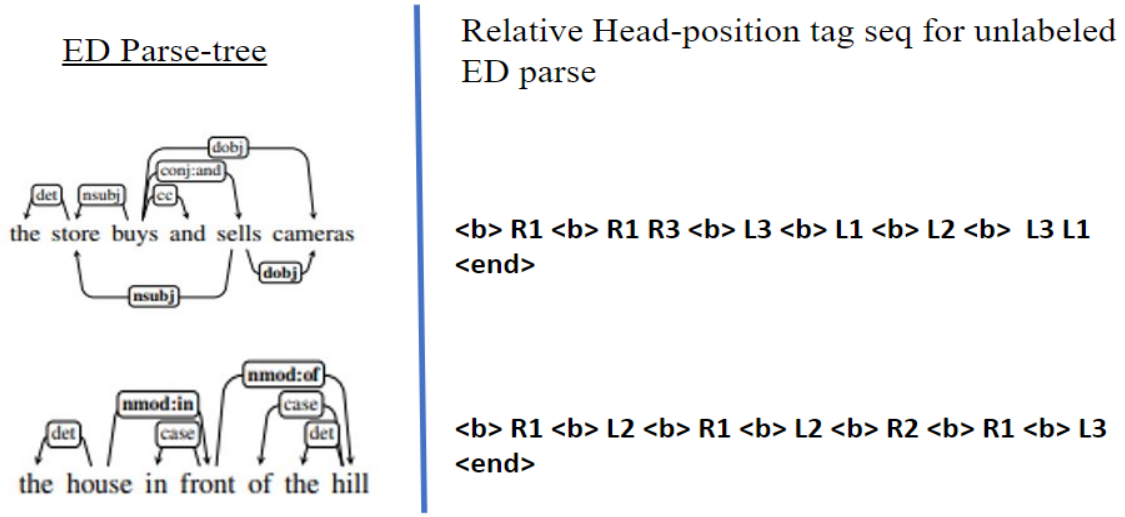

Figure 2: Example Enhanced Dependency Parse trees represented as *Relative Head-position tag-sequences*

spective relative head-position tag-sequences. Subsequently, the approach trains a standard LSTM-based model to predict the relative head-position tag for each token within an input-sentence. Results outlined in the paper show that this end-to-end parser performs as well as the state-of-the-art deep biaffine network (Dozat and Manning, 2016) while being much simpler in design.

2.2 UDify

UDify is an mBERT based multilingual model which simultaneously performs four key language-processing tasks; these tasks are *UPOS-tagging*, *UFeat-tagging*, *Lemmatization* and *Dependency Parsing*, in a multitasking framework. The model utilizes a single shared mBERT based encoder, and four individual task-specific decoders, for each of the four tasks respectively.

The *mBERT Encoder* takes in the entire sentence as input, tokenizes it using pre-trained the Word-Piece Tokenizer (Wu et al., 2016) and subsequently outputs mBERT (Wu and Dredze, 2019) based contextualized-embeddings for each word within the input-sentence. We refer to original UDify (Kondratyuk and Straka, 2019) paper for a detailed description of the mechanism of computing/fine-tuning such contextualized embeddings.

The decoders for both the *UPOS-tagging* and *UFeat-tagging* tasks adopt a standard sequence-tagging architecture with a softmax layer on the top. These decoders accept the contextual embeddings generated from the mBERT Encoder for each word in the input sentence, and predicts its UPOS/Ufeats tag.

For the *Lemmatization* task as well, the model uses a standard sequence-tagger which predicts a class-tag representing a unique edit script, for each word. An edit-script is simply the sequence of character operations to transform a word form to its lemma-form.

For dependency-parsing, the model adopts the popular deep biaffine architecture (Dozat and Manning, 2016) for graph-based parsing, with LSTM-encoder been replaced by the shared *mBERT En-*

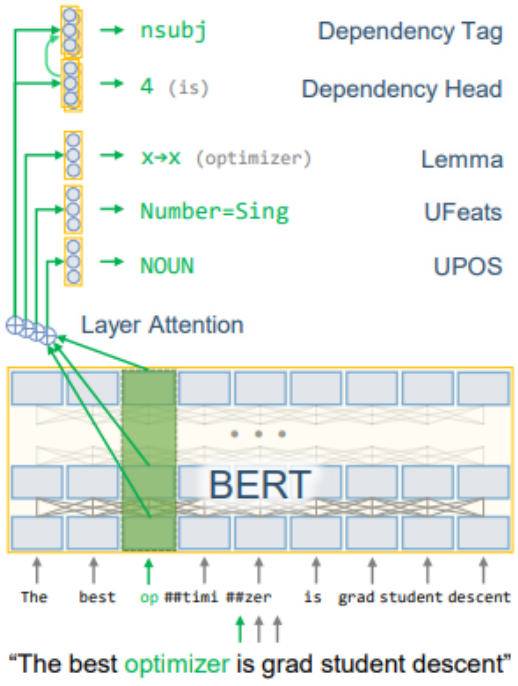

Figure 3: UDify model architecture. Figure from (Kondratyuk and Straka, 2019).

coder.

Hyper-parameter	Value
Dropout prob.	0.01
Bach-size	32
Number of steps per epoch	Size of training corpus / 32
Epochs	150
BERT Model	bert_multi_cased_L-12_H-768_A-12

Table 1: Hyper-parameters

3 mBERT based Seq2seq ED Parser

Figure 2b depicts the architecture of the proposed ED parser. Our proposed *End-to-end ED Parser* is an extension of the UDify (Kondratyuk and Straka, 2019) model described in section 2.2, with one additional component namely the *Relative Head Sequence predictor* which predicts the relative head-position of the tag-sequence representing the unlabelled enhanced-dependency parse-tree of the input sentence (as the fifth auxiliary task in the multitasking UDify model).

3.1 ED parse-tree as relative head-position tag sequence

Given a sentence of length T, its unlabelled ED parse-tree can be represented by a relative-head tag-seq of length $\hat{T}$ such that $\hat{T} \geq 2T + 1$. Figure 2 depicts the representations of sample unlabelled enhanced-dependency parse-trees as their relative sequences of relative head-position tags. Here, the tag $< b >$ represents the next-token whose heads are pointed by the subsequently predicted relative-head position tags (until the next $< b >$ tag is predicted).

3.2 Relative Head Sequence predictor

As evident in Figure 2b, our *Relative Head Sequence predictor* is a standard LSTM based Seq2seq neural-network (Sutskever et al., 2014) which takes in the entire input-sentence encoding vector as input, and sequentially predicts the relative head-position tag-sequence, one tag at a time.

3.2.1 Input sentence-encoding

The sentence-encoding $e^X \in R^d$ of any input sentence $X = x_1, x_2, ...x_T$ is computed by applying equation 1.

$$e^X = W * [BERT(X); TY_l] + b \qquad (1)$$

Here $BERT(X)$ is the output embedding-vector from the UDify's shared mBERT encoder for the end-of-sentence token $< /s >$ of input-sentence and TY_l is a *Linguistic-typology* vector of language l being parsed. Each value within TY_l represents a single typology-feature from *WALS* (Haspelmath, 2009) database having a specific integer value. Equation 1 involves the concatenation of the *BERT-output* and the *Typology* vectors, followed by dimension reduction through a feed-forward network. Feeding typology features together with the input sentence could improve the cross-lingual transferring ability of the multilingual model, as shown by (Ammar et al., 2016).

For the proposed model, we use all the word-order and constituency features in WALS (Haspelmath, 2009) database excluding trivially redundant features as excluded by (Takamura et al., 2016).

3.2.2 Training

We trained our *mBERT based Seq2seq ED Parser* on a single large joint-polyglot corpus, created by concatenating all the treebanks available in the training dataset provided for the *IWPT 2021 Shared task*.

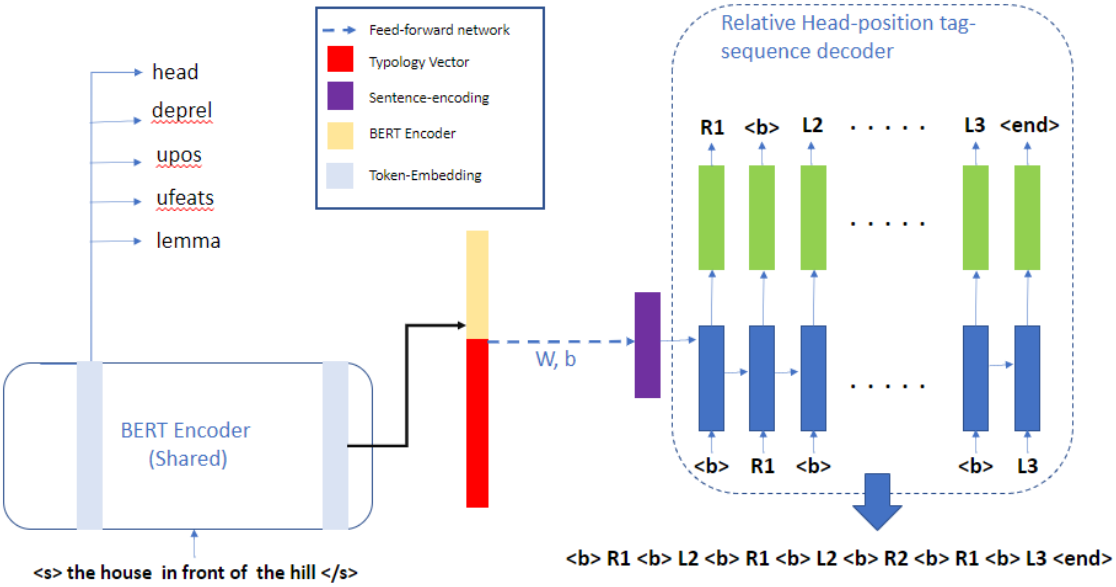

Figure 4: Architecture of the *Relative Head-position Sequence predictor*

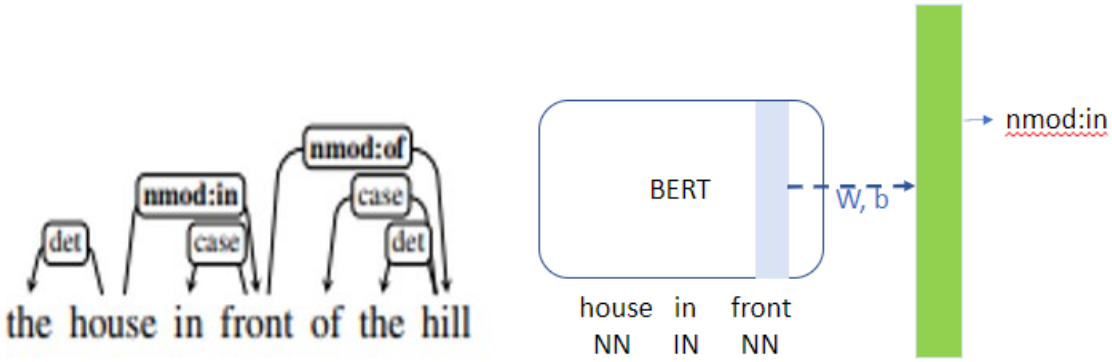

Figure 5: Architecture of the *Label predictor*

Before each training epoch, we randomly shuffle all sentences in our polyglot training corpus, and subsequently feed mixed batches of sentences from this shuffled corpus into the model being trained, where each batch may contain sentences from any language or treebank (as done by authors of UDify (Kondratyuk and Straka, 2019)).

We optimized the weights of our multitasking model by minimizing the total loss as the sum of sparse cross-entropy losses for all five tasks namely *UPOS-tagging*, *UFeat-tagging*, *Lemmatization*, *Dependency Parsing* and *Relative Head-position Sequence prediction*.

3.2.3 Predicting

The ED parsing of any unknown input-sentence $X = x_1, x_2, ...x_T$ can be performed by extracting the most probable correct relative head-position tag-sequence. The correct relative head-position tag-sequence would satisfy following constraints.

1. Sequence should start with $< b >$ and end with $< end >$.

2. For each word in $x_i \in X$, the relative head-position tag assigned to it should be within the range of the sentence. For example, within the sentence **"the house in front of the hill"**, the word '*the*' can not have tags L_2, L_3, L_4, L_5, L_6 and the word '*hill*' can not have any right tags, as these are outside the range of the sentence.

3. The label sequence should not generate any cycles within the dependency tree.

4. One of the words should have the head at $< root >$ token.

Language	UPOS	UFeats	Lemmas	UAS	LAS	ELAS
Bulgarian	98.81	35.97	97.40	93.37	90.03	78.45
English	95.17	32.77	95.76	87.07	84.46	65.40
Estonian	96.49	35.04	95.55	85.41	82.46	54.03
Latvian	96.12	35.61	95.45	88.51	85.19	56.67
Lithuanian	93.40	30.09	92.66	78.25	73.52	59.13
Russian	98.25	36.32	97.49	92.67	91.01	66.33
Slovak	96.62	22.68	94.61	90.09	87.49	67.45
Swedish	96.05	33.56	92.46	85.64	82.18	63.12

Table 2: Results achieved by our proposed End-to-end ED-parser

Model	UPOS	UFeats	Lemmas	UAS	LAS	EULAS	ELAS
combo	97.62	94.95	94.39	91.55	89.14	86.41	85.01
dcu-epfl	96.32	91.81	95.15	87.44	84.3	87.67	86.89
fastparse	97.24	93.0	95.84	78.23	72.44	69.42	67.07
grew	97.24	93.0	95.84	89.6	87.03	85.18	82.95
robertnlp	97.89	94.06	0.01	93.15	90.4	89.25	88.44
shanghaitech	0.46	32.78	0.01	4.18	1.27	89.76	88.37
tgif	0.46	32.81	0.01	10.93	0.94	91.45	90.67
unipi	96.37	91.75	95.17	90.55	87.98	85.96	84.42
nuig	**96.36**	**32.75**	**95.17**	**87.63**	**84.54**	**67.21**	**63.82**

Table 3: Average results achieved by all ED parsers

5. The sequence should contain the number of $< b >$ tags equal to number of tokens in the input sentence X.

We used dynamic programming with beam-search to efficiently extract the most probable relative head-position tag-sequence which satisfies the above listed relative head-position tag-sequence, out of all possible sequences.

3.3 Label Predictor

Figure 2c depicts the architecture of our *Label predictor* model. It is an mBERT based multi-class classifier with a softmax layer on top. The model takes as input the token-seq segment from the input sentence ranging from head to tail, as well as its corresponding predicted POS-tag sequence. The model outputs the probabilities of all possible ED dependency labels to be assigned to the given relation.

The *Label-predictor* is trained on all ED relationships available in training dataset for *IWPT 2021 Shared task*. The parameters of the mBERT encoder of our *Label predictor* are initialized with the parameters of the fine-tuned mBERT encoder of our *Relative Head-position tag-sequences*.

4 Experiments

As already explained, our proposed *End-to-end Seq2seq ED-parser* is trained on a large joint polyglot corpus created by concatenating all the treebanks in the provided training dataset for *IWPT 2021 Shared Task*. We evaluated our parser on test corpora provided for the *IWPT 2021 Shared Task* in eight distinct languages namely *Bulgarian*, *Estonian*, *English*, *Latvian*, *Lithuanian*, *Russian*, *Slovak* and *Swedish*. We outline the results achieved by our proposed model in detail in Section 5. Table 1 outlines hyper-parameters used in the experiments. These values are obtained by minimizing the training loss for *English-EWT Corpus* provided in the *dev* dataset provided for *IWPT 2021 Shared Task*.

5 Results and Conclusion

Table 2 outlines results achieved by our proposed *End-to-end BERT Based Se2seq ED-Parser* on all eight blind test-corpora on which the model is evaluated, as calculated by the evaluation script for the shared task.

Appendix A compares the results achieved by our *ED-parser* with the results achieved by the other participants of *IWPT 2021 Shared tasks*. Table 3 outlines the average results achieved by all

the models proposed in *IWPT 2021 Shared task* for all eight test-languages. It is evident that our models performs on par with other state-of-the-art ED-parsers despite the fact that its much simpler in design as it is an end-to-end design, and thus is much easier to train and implement.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2021. From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021. In *Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021)*, pages 146–157, Bangkok, Thailand (online). Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Martin Haspelmath. 2009. *The typological database of the World Atlas of Language Structures*. Berlin: Walter de Gruyter.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036, Paris, France. European Language Resources Association.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378.

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.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 69–76.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A Results

This section compares the results achieved by our *ED-parser* with the results achieved by the other participants of *IWPT 2021 Shared tasks*.

Table 4: Results of all participants of *IWPT 2021 Shared Task*

Begin of Table							
Language	**Models**	**UPOS**	**UFeats**	**Lemma**	**UAS**	**LAS**	**ELAS**
Bulgarian	combo	98.72	97.23	97.25	92.98	89.52	86.67
	dcu-epfl	98.89	97.57	97.30	93.25	90.19	92.44
	fastparse	99.15	97.95	97.97	87.85	83.39	78.73
	grew	99.15	97.95	97.97	94.36	91.62	88.83
	robertnlp	99.13	98.31	0.01	96.30	94.15	93.16
	shanghaitech	0.00	35.92	0.01	5.80	1.54	92.52
	tgif	0.00	35.98	0.01	10.58	1.13	93.63
	unipi	98.81	97.57	97.40	95.29	92.71	90.84
	nuig	**98.81**	**35.97**	**97.40**	**93.37**	**90.03**	**78.45**
English	combo	95.74	93.54	95.26	89.61	87.22	84.09
	dcu-epfl	94.96	93.53	95.66	86.45	83.64	85.70
	fastparse	95.85	94.16	96.04	82.36	77.99	73.00
	grew	95.85	94.16	96.04	89.22	86.83	85.49
	robertnlp	96.24	94.44	0.00	90.79	88.48	87.88
	shanghaitech	0.28	32.80	0.00	3.71	1.24	87.27
	tgif	0.28	32.76	0.00	7.86	1.08	88.19
	unipi	95.17	93.70	95.76	90.64	88.47	87.11
	nuig	**95.17**	**32.77**	**95.76**	**87.07**	**84.46**	**65.40**
Estonian	combo	97.42	96.57	86.09	90.00	87.53	84.02
	dcu-epfl	96.46	95.30	95.58	85.31	82.35	84.35
	fastparse	96.89	95.78	94.90	71.70	64.50	60.05
	grew	96.89	95.78	94.90	86.62	83.85	78.19
	robertnlp	97.09	96.46	0.00	90.02	87.59	86.55
	shanghaitech	0.12	34.99	0.00	3.67	1.16	86.66
	tgif	0.12	35.08	0.01	11.86	0.82	88.38
	unipi	96.49	95.33	95.55	87.11	84.14	81.27
	nuig	**96.49**	**35.04**	**95.55**	**85.41**	**82.46**	**54.03**
Latvian	combo	97.35	94.97	96.53	92.91	90.25	84.57
	dcu-epfl	95.95	93.59	95.34	88.47	85.10	86.96
	fastparse	96.28	93.79	95.81	78.37	72.03	66.43
	grew	96.28	93.79	95.81	88.32	85.27	77.45
	robertnlp	97.61	95.18	0.03	93.62	91.25	88.82
	shanghaitech	0.58	35.57	0.03	4.22	1.42	89.17
	tgif	0.56	35.62	0.03	10.37	0.97	90.23
	unipi	96.12	93.45	95.45	89.90	86.63	83.01
	nuig	**96.12**	**35.61**	**95.45**	**88.51**	**85.19**	**56.67**
Lithuanian	combo	97.26	95.05	93.76	88.03	84.75	79.75
	dcu-epfl	93.47	87.74	92.71	78.36	73.25	78.04
	fastparse	95.97	91.07	93.61	61.39	53.55	48.27
	grew	95.97	91.07	93.61	82.54	78.65	74.62
	robertnlp	97.42	93.20	0.00	90.49	83.27	80.76

	shanghaitech	1.51	30.12	0.00	5.12	1.77	80.87
	tgif	1.51	30.20	0.00	10.89	1.24	86.06
	unipi	93.40	87.14	92.66	82.75	78.31	71.31
	nuig	**93.40**	**30.09**	**92.66**	**78.25**	**73.52**	**59.13**
Russian	combo	98.94	98.04	98.16	95.37	94.29	90.73
	dcu-epfl	98.19	87.67	97.39	92.61	90.97	92.83
	fastparse	98.86	88.97	98.33	87.09	83.23	78.56
	grew	98.86	88.97	98.33	94.22	92.97	90.56
	robertnlp	99.06	89.51	0.00	95.65	94.64	92.64
	shanghaitech	0.02	36.35	0.00	3.35	0.73	93.59
	tgif	0.02	36.37	0.00	13.81	0.51	94.01
	unipi	98.25	87.52	97.49	94.51	93.32	90.90
	nuig	**98.25**	**36.32**	**97.49**	**92.67**	**91.01**	**66.33**
Slovak	combo	97.88	95.03	95.61	93.19	91.72	87.04
	dcu-epfl	96.55	91.15	94.72	89.27	86.60	89.59
	fastparse	97.67	93.42	96.47	78.23	71.71	64.28
	grew	97.67	93.42	96.47	92.27	90.45	86.92
	robertnlp	98.28	95.54	0.00	96.16	93.88	89.66
	shanghaitech	1.19	22.69	0.00	6.06	1.96	90.25
	tgif	1.17	22.69	0.00	13.67	1.60	94.96
	unipi	96.62	91.44	94.61	93.32	91.75	86.05
	nuig	**96.62**	**22.68**	**94.61**	**90.09**	**87.49**	**67.45**
Swedish	combo	97.67	89.19	92.45	90.31	87.82	83.20
	dcu-epfl	96.12	87.92	92.47	85.83	82.30	85.20
	fastparse	97.25	88.82	93.60	78.88	73.11	67.26
	grew	97.25	88.82	93.60	89.26	86.59	81.54
	robertnlp	98.30	89.87	0.00	92.15	89.92	88.03
	shanghaitech	0.00	33.79	0.00	1.55	0.34	86.62
	tgif	0.00	33.79	0.00	8.42	0.20	89.90
	unipi	96.07	87.83	92.47	90.86	88.53	84.91
	nuig	**96.05**	**33.56**	**92.46**	**85.64**	**82.18**	**63.12**
Language	**Models**	**UPOS**	**UFeats**	**Lemmas**	**UAS**	**LAS**	**ELAS**