Cross-Lingual Lemmatization and Morphology Tagging with Two-Stage Multilingual BERT Fine-Tuning

Daniel Kondratyuk
Charles University
Institute of Formal and Applied Linguistics
Saarland University
Department of Computational Linguistics
dankondratyuk@gmail.com

Abstract

We present our CHARLES-SAARLAND system for the SIGMORPHON 2019 Shared Task on Crosslinguality and Context in Morphology, in task 2, Morphological Analysis and Lemmatization in Context. We leverage the multilingual BERT model and apply several fine-tuning strategies introduced by UDify demonstrating exceptional evaluation performance on morpho-syntactic tasks. Our results show that fine-tuning multilingual BERT on the concatenation of all available treebanks allows the model to learn cross-lingual information that is able to boost lemmatization and morphology tagging accuracy over fine-tuning it purely monolingually. Unlike UDify, however, we show that when paired with additional character-level and word-level LSTM layers, a second stage of fine-tuning on each treebank individually can improve evaluation even further. Out of all submissions for this shared task, our system achieves the highest average accuracy and f1 score in morphology tagging and places second in average lemmatization accuracy.

1 Introduction

We focus on track 2 of the SIGMORPHON 2019 Shared Task (McCarthy et al., 2019), which requires systems to predict lemmas and morphosyntactic descriptions (MSDs) of words given sentences of pre-tokenized words. The data relies on treebanks provided by the Universal Dependencies (UD) project (Nivre et al., 2018), where MSDs are converted from UD format to the UniMorph schema (McCarthy et al., 2018; Kirov et al., 2018). Systems must predict from sentences given test data provided in 107 separate treebanks each representing one of 66 different languages.

Recent advances in contextual word representations show that pretraining language models on a large corpus of unsupervised text can be used to transfer their internal knowledge representations to other NLP tasks to boost evaluation scores significantly (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2018). We utilize the BERT base multilingual cased model pretrained on raw sentences found in the top 104 most-resourced languages of Wikipedia (Devlin et al., 2018) for all of our experiments. In addition, we use methods introduced by UDify (Kondratyuk, 2019) to

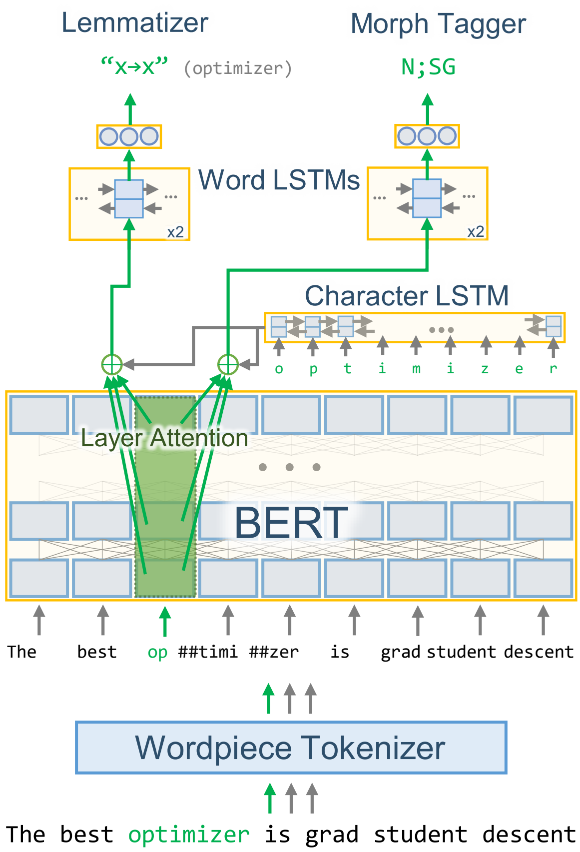

Figure 1: An illustration of our model architecture with task-specific layer attention, inputting word tokens and predicting lemma edit scripts and morphology tags for each token.

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 12–18
Florence, Italy. August 2, 2019

further fine-tune and regularize BERT, which has been shown to be especially helpful in predicting morpho-syntactic tasks.

Our system defines a simple multi-task multilingual neural architecture for predicting lemmas and MSDs jointly. Our contributions to achieve high lemmatization and morphology tagging performance are as follows:

1. We leverage the pretrained multilingual BERT cased model to encode input sentences and apply additional word-level and character-level LSTM layers before jointly decoding lemmas and morphology tags using simple sequence tagging layers.

2. Instead of only training models for each treebank separately, we use a two-stage training process to incorporate cross-linguistic information present in other treebanks, training multilingually over all treebanks in the first stage and then monolingually using saved multilingual weights in the second stage.

Our results show that applying an intermediate multilingual fine-tuning stage on BERT is superior to just fine-tuning monolingually in nearly all cases. Code for our model is released along with UDify at `https://github.com/hyperparticle/udify`.

2 Model Architecture

We describe the architecture of our system as follows. See Figure 1 for an illustration of this description. Our network consists of a shared BERT encoder followed by joint lemma and morphology tag decoders.

Given an input sentence consisting of a sequence of word tokens, we apply BERT's multilingual cased tokenizer to each word, potentially splitting it into multiple subword tokens. We encode this token sequence with the pretrained multilingual cased BERT base model consisting of 12 layers with 12 attention heads per layer and hidden output dimensions of 768. Following this, we take the subset of wordpieces corresponding to the first wordpiece of each word to align the BERT encoding with the sequence of input words[1].

Once BERT encoding is complete, we apply two separate instances of layer attention defined in UDify which is similar to ELMo (Peters et al., 2018), i.e., a trainable weighted sum of all 12 layers of BERT, which has been shown to improve evaluation performance over just computing representations on the last layer. The layer attention instances generate embeddings specific to each task, one for lemmatization and the other for morphology tagging.

But before decoding, we also apply character-level embeddings (Santos and Zadrozny, 2014; Ling et al., 2015; Kim et al., 2016) to produce an enhanced morphological representation by encoding the sequence of character tokens for each word through a bidirectional LSTM with a residual connection (Schuster and Paliwal, 1997; Kim et al., 2017), keeping the hidden layers fixed to dimensions of 384. We concatenate the final hidden states of both LSTM directions, and then sum these character-level word representations with each of the two encoded representations produced by the task-specific layer attention.

Similar to Kondratyuk et al. (2018) and Straka (2018), both the lemmatizer and morphological tagger employ two successive layers of word-level bidirectional residual LSTMs computed over the entire task layer attention sequence with hidden dimensions of 768, summing both directions together along each output state.

For lemmatization, we precompute edit scripts representing a minimal sequence of character operations to transduce a word form to its lemma counterpart, as seen in Chrupała (2006); Straka (2018). As is typical for neural sequence tagging, we apply a feedforward layer to the final layer of the lemmatizer LSTM, representing the activations of classes of all edit scripts found in the training data.

Similarly for morphology tagging, we apply a feedforward layer whose units correspond to the vocabulary over all unfactored MSD strings. We apply the method of Inoue et al. (2017) to jointly predict the classes of unfactored and factored morphology tags, i.e., we also predict each dimension of the morphology tag whose subcategories are defined by the UniMorph schema (e.g., case, mood, person, tense, etc.). We only use the factored tags to improve training, and for prediction we use the full unfactored tags.

[1] Kondratyuk (2019) and Kitaev and Klein (2018) found that first, last, or average of the wordpieces did not make a noticeable difference.

HYPERPARAMETER	VALUE
Character-level embedding dimension	256
Character-level LSTM hidden dimension	384
Word-level LSTM hidden dimension	768
Final feedforward learning rate	$4e^{-3}$
LSTM, layer attention learning rate	$1e^{-3}$
BERT learning rate (layers 7-12)	$5e^{-5}$
BERT learning rate (layers 1-6)	$1e^{-5}$
LSTM embedding dropout	0.5
BERT internal dropout	0.25
Mask probability	0.25
Layer dropout	0.2
Batch size	32
Epochs	50

Table 1: A summary of hyperparameters applicable to each model configuration.

3 Experiments

We train our system on the provided treebank training data with three separate configurations.

3.1 Configurations

MONO We train the network (as seen in Figure 1) monolingually by simply fine-tuning it on each treebank separately.

MULTI We fine-tune the network as in MONO, except on a dataset consisting of all treebank training data concatenated together, as seen in UDify. All word, character, and tag vocabularies of each language are combined together.

MULTI+MONO We train the network monolingually as in MONO, but using the BERT weights saved from the model fine-tuned according to MULTI. This effectively defines a two-stage training process: the first stage involves multilingual fine-tuning of BERT, and the second stage re-trains the layer attention, LSTMs and feedforward taggers from scratch on each treebank with a reduced monolingual vocabulary (keeping fine-tuned BERT intact).

For all MONO and the second stage of MULTI+MONO, we ensure that we do not combine multiple treebanks of the same language but always fine-tune on just the training data from each provided treebank.

3.2 Hyperparameters

A summary of specific values for each of the hyperparameters discussed can be seen in Table 1.

We train each configuration using a batch size of 32 over 50 epochs. We employ the Adam optimizer, computing the loss as the softmax cross entropy between the predicted tags and the gold labels. We apply discriminative fine-tuning (Howard and Ruder, 2018) by defining four separate parameter groups each with their own base learning rate, decreasing as the layers get closer to the input: the first 6 layers of BERT, the last 6 layers of BERT, the layer attention and LSTM layers, and the final feedforward layers.

We apply regularization as defined by UDify, with a few extra modifications. We raise the layer dropout, BERT dropout, input mask probability slightly to prevent overfitting, especially for the MONO and MULTI+MONO configurations. We also apply dropout to all intermediate word-embedding representations between each of the word-level LSTM layers.

MODEL	LEMMA ACC	LEMMA DIST	MORPH ACC	MORPH F1
Baseline	93.13	0.13	73.16	87.92
Mono	92.80	0.17	90.26	93.44
Multi	90.39	0.27	85.18	90.18
Multi+Mono	**95.00**	**0.12**	**93.23**	**96.02**

Table 2: A summary of the average results of each model configuration with a comparison to the baseline (Malaviya et al., 2019).

4 Results

We display comparisons between each of the three configurations. We compute lemma accuracy, lemma Levenstein distance, morphology tag accuracy, and morphology f1 scores for each of the 107 treebanks. A summary of the averages of all scores for each configuration can be found in Table 2. The full results are shown in Tables 3, 4, 5, and 6.

5 Discussion

Our results show that not only does fine-tuning BERT provide excellent lemmatization and morphology tagging performance, two-stage MULTI+MONO training can provide significant improvements for practically every treebank when compared to MONO. While some of these improvements can be attributed to learning from monolingual data from multiple treebanks of the same language, we can see improvements even for languages possessing just one treebank. This provides evidence that the MULTI and MULTI+MONO models regularize well to multilingual training. This could be explained by a combination of: multilingual learning providing

Treebank	Model	Lemma		Morph	
		Acc	Dist	Acc	F1
Afrikaans AfriBooms	Mono	98.66	0.03	98.4	98.63
	Multi	97.19	0.05	98.06	98.58
	Multi+Mono	**98.95**	**0.02**	**99.23**	**99.36**
Akkadian PISANDUB	Mono	49.78	2.14	86.22	86.41
	Multi	23.56	3.63	60.44	60.89
	Multi+Mono	**65.35**	**0.97**	**89.11**	**89.06**
Amharic ATT	Mono	**100.0**	**0.00**	86.8	90.74
	Multi	**100.0**	**0.00**	81.0	86.14
	Multi+Mono	**100.0**	**0.00**	**87.43**	**91.34**
Ancient Greek PROIEL	Mono	92.15	0.22	90.85	96.95
	Multi	85.75	0.43	88.99	96.2
	Multi+Mono	**92.34**	**0.20**	**92.37**	**97.68**
Ancient Greek Perseus	Mono	88.88	0.32	88.9	94.74
	Multi	80.98	0.56	86.38	93.4
	Multi+Mono	**89.69**	**0.29**	**90.88**	**96.26**
Arabic PADT	Mono	94.24	0.17	94.09	96.91
	Multi	75.54	0.85	93.66	96.88
	Multi+Mono	**94.45**	**0.16**	**95.66**	**97.65**
Arabic PUD	Mono	71.0	1.50	84.03	93.78
	Multi	36.65	5.03	65.25	85.59
	Multi+Mono	**81.92**	**0.48**	**84.53**	**94.09**
Armenian ArmTDP	Mono	94.5	0.10	91.05	95.48
	Multi	91.48	0.17	82.49	89.92
	Multi+Mono	**95.58**	**0.08**	**92.77**	**96.66**
Bambara CRB	Mono	**90.08**	**0.18**	92.7	94.02
	Multi	72.25	0.58	77.09	81.81
	Multi+Mono	88.76	0.21	**93.32**	**95.34**
Basque BDT	Mono	96.3	0.08	90.03	94.72
	Multi	93.72	0.14	85.54	92.76
	Multi+Mono	**96.5**	**0.07**	**92.07**	**96.3**
Belarusian HSE	Mono	87.76	0.21	78.62	89.47
	Multi	87.62	0.22	73.56	81.76
	Multi+Mono	**92.51**	**0.12**	**89.93**	**95.68**
Breton KEB	Mono	**91.19**	**0.20**	**90.88**	**92.93**
	Multi	80.24	0.55	76.49	79.07
	Multi+Mono	87.66	0.32	90.35	91.77
Bulgarian BTB	Mono	96.72	0.10	96.61	98.3
	Multi	95.06	0.16	95.64	98.02
	Multi+Mono	**98.05**	**0.07**	**98.01**	**99.18**
Buryat BDT	Mono	85.48	0.33	80.29	82.5
	Multi	73.48	0.57	64.25	67.12
	Multi+Mono	**86.35**	**0.30**	**85.67**	**88.42**
Cantonese HK	Mono	99.49	0.01	92.11	90.19
	Multi	98.63	0.02	87.31	84.65
	Multi+Mono	**100.0**	**0.00**	**94.29**	**92.83**
Catalan AnCora	Mono	99.2	**0.01**	98.36	99.19
	Multi	98.87	0.02	98.58	99.37
	Multi+Mono	**99.38**	**0.01**	**98.82**	**99.45**
Chinese CFL	Mono	**100.0**	**0.00**	92.52	91.46
	Multi	**100.0**	**0.00**	84.9	85.56
	Multi+Mono	99.65	**0.00**	**92.55**	**91.5**
Chinese GSD	Mono	99.94	**0.00**	94.56	94.44
	Multi	**100.0**	**0.00**	97.03	96.96
	Multi+Mono	99.97	**0.00**	**97.13**	**97.04**
Coptic Scriptorium	Mono	92.52	0.17	89.93	92.28
	Multi	84.75	0.33	78.69	82.32
	Multi+Mono	**96.13**	**0.08**	**93.3**	**94.81**
Croatian SET	Mono	96.73	0.06	92.07	96.86
	Multi	96.54	0.06	91.01	96.74
	Multi+Mono	**97.51**	**0.05**	**94.11**	**97.82**
Czech CAC	Mono	99.03	0.02	96.43	98.67
	Multi	99.04	0.02	97.09	99.07
	Multi+Mono	**99.45**	**0.01**	**98.48**	**99.48**
Czech CLTT	Mono	98.09	0.03	92.35	96.63
	Multi	99.29	**0.01**	92.99	97.49
	Multi+Mono	**99.3**	**0.01**	**95.31**	**98.2**
Czech FicTree	Mono	98.11	0.03	93.39	97.14
	Multi	98.62	0.03	92.06	97.39
	Multi+Mono	**99.01**	**0.02**	**97.13**	**98.9**
Czech PDT	Mono	99.14	**0.01**	97.01	98.84
	Multi	99.12	0.02	97.48	99.12
	Multi+Mono	**99.42**	**0.01**	**98.54**	**99.47**
Czech PUD	Mono	92.71	0.12	80.71	92.13
	Multi	**97.91**	**0.03**	**92.71**	**97.64**
	Multi+Mono	96.74	0.06	92.38	97.43
Danish DDT	Mono	96.48	0.06	95.72	97.15
	Multi	96.47	0.07	96.25	97.73
	Multi+Mono	**98.15**	**0.03**	**97.98**	**98.68**
Dutch Alpino	Mono	97.63	0.04	96.64	97.43
	Multi	96.71	0.07	97.51	98.24
	Multi+Mono	**98.62**	**0.03**	**98.12**	**98.62**
Dutch LassySmall	Mono	96.77	0.06	96.11	97.0
	Multi	97.41	0.06	98.04	98.6
	Multi+Mono	**98.08**	**0.03**	**98.5**	**98.83**

Table 3: Main results (part 1 of 4).

Treebank	Model	Lemma		Morph	
		Acc	Dist	Acc	F1
English EWT	Mono	98.56	0.02	96.44	97.38
	Multi	98.49	0.03	96.98	97.99
	Multi+Mono	**99.19**	**0.01**	**97.85**	**98.52**
English GUM	Mono	97.75	0.04	96.17	97.11
	Multi	94.97	0.09	93.6	96.15
	Multi+Mono	**98.45**	**0.02**	**97.52**	**98.11**
English LinES	Mono	98.31	0.03	96.76	97.51
	Multi	96.6	0.07	93.06	95.48
	Multi+Mono	**98.62**	**0.02**	**97.77**	**98.3**
English PUD	Mono	95.98	0.06	95.89	97.0
	Multi	94.05	0.13	92.65	95.76
	Multi+Mono	**97.89**	**0.03**	**96.67**	**97.58**
English ParTUT	Mono	97.87	0.03	96.02	96.55
	Multi	97.8	0.04	92.72	94.98
	Multi+Mono	**98.51**	**0.02**	**96.65**	**97.35**
Estonian EDT	Mono	**93.21**	**0.15**	95.3	97.56
	Multi	89.13	0.23	96.13	98.18
	Multi+Mono	88.16	0.22	**97.23**	**98.69**
Faroese OFT	Mono	**89.14**	0.22	**86.97**	92.27
	Multi	79.94	0.40	75.8	83.55
	Multi+Mono	88.95	**0.20**	86.74	**93.47**
Finnish FTB	Mono	93.74	0.13	93.61	96.3
	Multi	93.25	0.12	92.67	96.75
	Multi+Mono	**95.45**	**0.08**	**96.85**	**98.38**
Finnish PUD	Mono	75.7	0.54	89.28	94.22
	Multi	**85.71**	**0.21**	94.76	97.69
	Multi+Mono	85.48	0.28	**95.62**	**97.98**
Finnish TDT	Mono	93.89	0.12	95.05	97.05
	Multi	92.76	0.13	93.41	97.1
	Multi+Mono	**95.73**	**0.08**	**97.1**	**98.54**
French GSD	Mono	98.51	0.03	97.51	98.57
	Multi	98.44	0.03	97.64	98.85
	Multi+Mono	**99.01**	**0.02**	**98.31**	**99.07**
French ParTUT	Mono	94.88	0.10	94.35	97.2
	Multi	94.1	0.13	91.56	96.74
	Multi+Mono	**96.66**	**0.06**	**95.46**	**97.95**
French Sequoia	Mono	97.86	0.04	96.57	98.2
	Multi	98.36	0.03	92.56	97.45
	Multi+Mono	**98.81**	**0.02**	**97.75**	**98.99**
French Spoken	Mono	97.67	0.04	98.07	98.09
	Multi	98.42	0.03	97.17	97.2
	Multi+Mono	**98.85**	**0.02**	**98.6**	**98.65**
Galician CTG	Mono	98.58	**0.02**	98.23	98.07
	Multi	98.19	0.03	96.94	96.47
	Multi+Mono	**98.96**	**0.02**	**98.44**	**98.29**
Galician TreeGal	Mono	95.32	0.07	92.14	95.32
	Multi	96.24	0.05	84.58	92.1
	Multi+Mono	**98.46**	**0.03**	**96.21**	**97.88**
German GSD	Mono	97.18	0.06	88.01	94.75
	Multi	95.89	0.10	89.33	95.46
	Multi+Mono	**97.62**	**0.05**	**90.43**	**95.9**
Gothic PROIEL	Mono	93.25	0.14	85.95	93.72
	Multi	86.86	0.29	86.06	94.16
	Multi+Mono	**94.54**	**0.13**	**91.02**	**96.64**
Greek GDT	Mono	**94.64**	**0.13**	92.74	97.21
	Multi	90.99	0.25	92.36	97.12
	Multi+Mono	82.95	0.42	**95.61**	**98.23**
Hebrew HTB	Mono	96.55	0.07	95.99	97.2
	Multi	95.25	0.09	95.45	97.3
	Multi+Mono	**97.85**	**0.04**	**97.67**	**98.47**
Hindi HDTB	Mono	98.7	0.02	91.95	97.16
	Multi	98.63	0.02	92.16	97.4
	Multi+Mono	**98.84**	**0.01**	**93.65**	**98.04**
Hungarian Szeged	Mono	93.68	0.12	84.9	94.42
	Multi	92.82	0.12	79.01	92.69
	Multi+Mono	**96.99**	**0.06**	**91.5**	**97.51**
Indonesian GSD	Mono	99.4	**0.01**	90.62	93.84
	Multi	98.92	0.02	90.84	94.01
	Multi+Mono	**99.51**	**0.01**	**92.48**	**95.16**
Irish IDT	Mono	**89.07**	**0.26**	80.44	86.01
	Multi	85.9	0.33	75.72	84.6
	Multi+Mono	88.09	0.27	**84.4**	**90.04**
Italian ISDT	Mono	98.33	0.03	97.88	98.77
	Multi	98.34	0.03	98.31	99.17
	Multi+Mono	**98.88**	**0.02**	**98.49**	**99.19**
Italian PUD	Mono	94.82	0.10	95.1	97.67
	Multi	96.19	0.09	59.07	84.9
	Multi+Mono	**97.69**	**0.04**	**96.37**	**98.42**
Italian ParTUT	Mono	97.32	0.05	97.32	98.24
	Multi	98.24	0.04	97.92	98.8
	Multi+Mono	**98.87**	**0.02**	**98.4**	**99.2**
Italian PoSTWITA	Mono	96.15	0.08	95.87	96.82
	Multi	95.24	0.12	96.56	97.58
	Multi+Mono	**97.24**	**0.06**	**96.88**	**97.9**

Table 4: Main results (part 2 of 4).

		Lemma		Morph	
Treebank	Model	Acc	Dist	Acc	F1
Japanese GSD	Mono	99.36	0.01	97.36	97.04
	Multi	99.49	0.01	98.07	97.83
	Multi+Mono	**99.65**	**0.00**	**98.41**	**98.21**
Japanese Modern	Mono	96.17	0.05	96.1	96.17
	Multi	94.57	0.08	90.05	90.16
	Multi+Mono	**98.67**	**0.01**	**97.47**	**97.5**
Japanese PUD	Mono	98.89	0.02	96.78	96.45
	Multi	**99.5**	**0.01**	97.9	97.7
	Multi+Mono	99.36	**0.01**	**98.56**	**98.39**
Komi Zyrian IKDP	Mono	56.63	0.88	45.78	49.74
	Multi	63.86	0.83	38.55	37.04
	Multi+Mono	**78.91**	**0.38**	**67.97**	**75.05**
Komi Zyrian Lattice	Mono	63.74	0.82	44.51	52.06
	Multi	60.44	1.05	39.56	45.87
	Multi+Mono	**80.77**	**0.36**	**67.58**	**78.01**
Korean GSD	Mono	87.47	0.26	96.18	95.66
	Multi	83.82	0.35	94.06	93.22
	Multi+Mono	**91.95**	**0.16**	**96.77**	**96.27**
Korean Kaist	Mono	92.62	0.14	96.97	96.59
	Multi	89.3	0.23	97.54	97.24
	Multi+Mono	**93.18**	**0.12**	**97.85**	**97.58**
Korean PUD	Mono	98.56	0.03	92.36	95.51
	Multi	68.19	0.99	64.7	70.71
	Multi+Mono	**99.57**	**0.01**	**94.67**	**96.76**
Kurmanji MG	Mono	87.54	0.24	80.69	86.67
	Multi	78.91	0.45	65.04	72.29
	Multi+Mono	**93.73**	**0.12**	**84.23**	**90.26**
Latin ITTB	Mono	98.68	0.03	95.17	97.65
	Multi	98.53	0.04	96.38	98.44
	Multi+Mono	**99.2**	**0.02**	**97.64**	**98.96**
Latin PROIEL	Mono	95.75	0.09	88.81	95.43
	Multi	94.67	0.12	91.15	96.78
	Multi+Mono	**97.36**	**0.05**	**93.68**	**97.87**
Latin Perseus	Mono	79.04	0.43	72.1	83.21
	Multi	86.43	0.27	80.53	90.8
	Multi+Mono	**89.68**	**0.19**	**85.94**	**93.79**
Latvian LVTB	Mono	95.15	0.08	92.59	95.85
	Multi	94.73	0.09	91.88	95.75
	Multi+Mono	**97.14**	**0.05**	**95.78**	**98.04**
Lithuanian HSE	Mono	74.46	0.53	67.6	75.01
	Multi	73.61	0.48	66.09	78.74
	Multi+Mono	**85.57**	**0.25**	**79.46**	**87.97**
Marathi UFAL	Mono	73.65	0.67	59.53	74.76
	Multi	75.53	0.65	55.53	75.05
	Multi+Mono	**76.69**	**0.61**	**67.75**	**80.19**
Naija NSC	Mono	99.84	0.01	95.64	94.16
	Multi	99.43	0.01	92.33	89.49
	Multi+Mono	**100.0**	**0.00**	**96.5**	**95.31**
North Sami Giella	Mono	85.74	0.30	84.66	90.44
	Multi	79.06	0.42	83.28	90.03
	Multi+Mono	**90.17**	**0.21**	**92.46**	**95.33**
Norwegian Bokmaal	Mono	98.76	0.02	97.13	98.32
	Multi	98.62	0.02	97.73	98.83
	Multi+Mono	**99.18**	**0.01**	**98.25**	**99.02**
Norwegian Nynorsk	Mono	98.45	0.02	96.89	98.17
	Multi	98.34	0.03	97.62	98.77
	Multi+Mono	**99.0**	**0.01**	**98.11**	**98.97**
Norwegian NynorskLIA	Mono	96.24	0.07	93.37	94.96
	Multi	97.28	0.05	93.96	96.29
	Multi+Mono	**98.08**	**0.04**	**96.8**	**97.39**
Old Church Slavonic PROIEL	Mono	91.09	0.19	86.24	93.22
	Multi	82.79	0.39	80.18	89.63
	Multi+Mono	**93.7**	**0.15**	**91.71**	**96.45**
Persian Seraji	Mono	95.34	0.23	97.17	97.69
	Multi	92.17	0.41	96.85	97.73
	Multi+Mono	**96.63**	**0.17**	**98.31**	**98.67**
Polish LFG	Mono	96.25	0.07	92.44	96.74
	Multi	96.01	0.09	89.63	96.47
	Multi+Mono	**97.94**	**0.04**	**97.13**	**98.86**
Polish SZ	Mono	96.54	0.07	88.15	94.79
	Multi	96.22	0.08	69.63	91.35
	Multi+Mono	**97.43**	**0.05**	**95.11**	**98.11**
Portuguese Bosque	Mono	98.26	0.03	95.1	97.57
	Multi	97.48	0.05	94.45	97.34
	Multi+Mono	**98.65**	**0.02**	**96.22**	**98.26**
Portuguese GSD	Mono	98.64	0.07	98.63	98.74
	Multi	97.73	0.11	98.05	98.03
	Multi+Mono	**99.09**	**0.05**	**99.03**	**99.1**
Romanian Nonstandard	Mono	95.66	0.08	93.15	96.26
	Multi	92.88	0.13	93.8	96.99
	Multi+Mono	**96.52**	**0.06**	**95.01**	**97.65**
Romanian RRT	Mono	97.98	0.03	97.34	98.19
	Multi	97.14	0.05	97.15	98.36
	Multi+Mono	**98.58**	**0.02**	**98.19**	**98.89**
Russian GSD	Mono	96.41	0.06	90.73	95.92
	Multi	97.34	**0.04**	90.6	96.58
	Multi+Mono	**97.74**	**0.04**	**94.92**	**97.95**

Table 5: Main results (part 3 of 4).

		Lemma		Morph	
Treebank	Model	Acc	Dist	Acc	F1
Russian PUD	Mono	89.44	0.19	86.15	93.84
	Multi	94.38	0.10	64.26	89.43
	Multi+Mono	**95.49**	**0.08**	**91.15**	**96.27**
Russian SynTagRus	Mono	98.6	0.03	97.22	98.61
	Multi	98.28	0.04	97.76	98.97
	Multi+Mono	**99.01**	**0.02**	**98.38**	**99.23**
Russian Taiga	Mono	88.91	0.20	82.64	88.88
	Multi	**94.13**	**0.13**	88.61	**94.89**
	Multi+Mono	93.49	**0.13**	**90.15**	94.88
Sanskrit UFAL	Mono	57.22	1.12	43.81	58.11
	Multi	49.48	1.24	33.51	43.14
	Multi+Mono	**63.32**	**0.89**	**47.74**	**69.52**
Serbian SET	Mono	96.74	0.06	93.86	97.02
	Multi	97.36	0.05	93.22	97.18
	Multi+Mono	**98.08**	**0.03**	**97.02**	**98.64**
Slovak SNK	Mono	96.31	0.06	89.24	95.15
	Multi	95.73	0.07	90.61	96.23
	Multi+Mono	**97.57**	**0.04**	**95.41**	**98.24**
Slovenian SSJ	Mono	97.22	0.04	92.56	96.37
	Multi	97.6	0.04	92.97	97.2
	Multi+Mono	**98.87**	**0.02**	**97.01**	**98.8**
Slovenian SST	Mono	93.46	0.10	83.46	90.38
	Multi	**97.24**	**0.05**	87.76	94.06
	Multi+Mono	97.2	**0.05**	**92.76**	**96.2**
Spanish AnCora	Mono	99.07	0.02	98.15	99.04
	Multi	98.87	0.02	98.36	99.19
	Multi+Mono	**99.4**	**0.01**	**98.79**	**99.4**
Spanish GSD	Mono	99.0	**0.01**	**95.93**	98.05
	Multi	98.35	0.02	95.63	97.96
	Multi+Mono	**99.16**	**0.01**	95.88	**98.08**
Swedish LinES	Mono	96.24	0.07	93.49	96.42
	Multi	94.94	0.09	92.43	96.6
	Multi+Mono	**97.83**	**0.04**	**94.75**	**97.67**
Swedish PUD	Mono	91.83	0.12	93.26	95.64
	Multi	90.81	0.14	93.46	96.37
	Multi+Mono	**95.85**	**0.07**	**95.62**	**97.25**
Swedish Talbanken	Mono	97.54	0.04	96.46	97.92
	Multi	97.17	0.05	96.65	98.52
	Multi+Mono	**98.62**	**0.02**	**98.09**	**99.05**
Tagalog TRG	Mono	76.0	0.48	72.0	79.17
	Multi	72.0	0.60	28.0	38.2
	Multi+Mono	**91.89**	**0.19**	**91.89**	**95.04**
Tamil TTB	Mono	88.96	0.27	84.78	92.41
	Multi	90.58	0.22	80.21	88.04
	Multi+Mono	**91.52**	**0.20**	**91.07**	**95.64**
Turkish IMST	Mono	93.43	**0.11**	85.51	91.71
	Multi	91.77	0.12	76.86	87.72
	Multi+Mono	**94.77**	**0.11**	**90.55**	**95.38**
Turkish PUD	Mono	83.11	0.37	84.34	92.13
	Multi	84.92	0.36	49.33	76.78
	Multi+Mono	**86.52**	**0.32**	**87.47**	**94.43**
Ukrainian IU	Mono	96.14	0.06	90.68	95.59
	Multi	96.31	0.06	91.42	96.32
	Multi+Mono	**97.84**	**0.03**	**95.78**	**98.1**
Upper Sorbian UFAL	Mono	85.25	0.25	74.19	81.49
	Multi	83.19	0.31	71.96	81.55
	Multi+Mono	**93.74**	**0.10**	**86.37**	**92.54**
Urdu UDTB	Mono	96.34	0.07	78.57	92.0
	Multi	96.08	0.07	79.26	92.44
	Multi+Mono	**96.92**	**0.06**	**80.67**	**93.45**
Vietnamese VTB	Mono	**99.81**	**0.00**	93.5	92.99
	Multi	99.35	0.01	93.96	93.47
	Multi+Mono	99.75	**0.00**	**94.54**	**94.02**
Yoruba YTB	Mono	97.6	**0.02**	88.0	85.33
	Multi	96.4	0.04	75.6	70.99
	Multi+Mono	**98.45**	**0.02**	**93.02**	**93.15**

Table 6: Main results (part 4 of 4).

language-invariant generalizations, out-of-domain data providing noise to reduce overfitting, or warm restarts aiding in improved convergence of model parameters. More experimentation is necessary to quantify these possible contributors.

Unlike the results shown by UDify, we see that the MULTI configuration provides overall inferior predictions on almost every treebank when compared to both MONO and MULTI+MONO.

This is likely due to the added LSTM layers and character-level embeddings, which provide additional information that improves monolingual training representations far more than it improves multilingual. Our intuition is that the LSTM layers pose an information bottleneck for massively multilingual data, unlike the BERT encoder, whose large capacity has been shown to be able to scale to more than 100 languages. Predictions using a smaller vocabulary subset could provide a much stronger signal to the LSTM layers to incorporate character-level morphology more accurately. But we do see that learning MULTI still learns useful cross-lingual information, just that it requires the LSTMs and character embeddings to be reconfigured to the specific treebank at hand to gain the benefits of both types of training.

Note that we specifically do not perform any extensive hyperparameter search or use ensembling. As such, we predict that our evaluation results could still be raised much higher.

6 Conclusion

We have demonstrated our system consisting of fine-tuning a multi-task enhanced BERT model for lemmatization and morphology tagging using a two-stage multilingual training scheme. We show that while pretrained BERT does provide word representations capable of surpassing the baseline, we are able to improve this significantly by also incorporating multilingual pretraining on all available treebanks, allowing the model to regularize and likely incorporate cross-lingual information useful for morphological parsing. We leave a more detailed analysis as to what extent multilingual fine-tuning and BERT pretraining contribute to model performance for future work.

7 Acknowledgements

Daniel Kondratyuk has been supported by the Erasmus Mundus program in Language & Communication Technologies (LCT).

References

Grzegorz Chrupała. 2006. Simple data-driven context-sensitive lemmatization. *Procesamiento del Lenguaje Natural*, 37.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339.

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 (CoNLL 2017)*, pages 421–431.

Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee. 2017. Residual lstm: Design of a deep recurrent architecture for distant speech recognition. *Proc. Interspeech 2017*, pages 1591–1595.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

Daniel Kondratyuk. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

Daniel Kondratyuk, Tomáš Gavenčiak, Milan Straka, and Jan Hajič. 2018. Lemmatag: Jointly tagging and lemmatizing for morphologically rich languages with brnns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4921–4928.

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.

Chaitanya Malaviya, Shijie Wu, and Ryan Cotterell. 2019. A simple joint model for improved contextual neural lemmatization. *arXiv preprint arXiv:1904.02306v2*.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Crosslinguality and context in morphology. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Florence, Italy. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, and Ahrenberg. 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 E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

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 (ICML-14)*, pages 1818–1826.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 45(11):2673.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.