Low-Resource Parsing with Crosslingual Contextualized Representations

Phoebe Mulcaire[♡*] **Jungo Kasai**[♡*] **Noah A. Smith**[♡◇]
[♡]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
[◇]Allen Institute for Artificial Intelligence, Seattle, WA, USA
{pmulc,jkasai,nasmith}@cs.washington.edu

Abstract

Despite advances in dependency parsing, languages with small treebanks still present challenges. We assess recent approaches to multilingual contextual word representations (CWRs), and compare them for crosslingual transfer from a language with a large treebank to a language with a small or nonexistent treebank, by sharing parameters between languages in the parser itself. We experiment with a diverse selection of languages in both simulated and truly low-resource scenarios, and show that multilingual CWRs greatly facilitate low-resource dependency parsing even without crosslingual supervision such as dictionaries or parallel text. Furthermore, we examine the non-contextual part of the learned language models (which we call a "decontextual probe") to demonstrate that polyglot language models better encode crosslingual lexical correspondence compared to aligned monolingual language models. This analysis provides further evidence that polyglot training is an effective approach to crosslingual transfer.

1 Introduction

Dependency parsing has achieved new states of the art using distributed word representations in neural networks, trained with large amounts of annotated data (Dozat and Manning, 2017; Dozat et al., 2017; Ma et al., 2018; Che et al., 2018). However, many languages are low-resource, with small or no treebanks, which presents a severe challenge in developing accurate parsing systems in those languages. One way to address this problem is with a crosslingual solution that makes use of a language with a large treebank and raw text in both languages. The hypothesis behind this approach is that, although each language is unique, different languages manifest similar characteristics (e.g., morphological, lexical, syntactic) which can be exploited by training a single *polyglot* model with data from multiple languages (Ammar, 2016).

Recent work has extended contextual word representations (CWRs) multilingually either by training a polyglot language model (LM) on a mixture of data from multiple languages (*joint training* approach; Mulcaire et al., 2019; Lample and Conneau, 2019) or by aligning multiple monolingual language models crosslingually (*retrofitting* approach; Schuster et al., 2019; Aldarmaki and Diab, 2019). These multilingual representations have been shown to facilitate crosslingual transfer on several tasks, including Universal Dependencies parsing and natural language inference. In this work, we assess these two types of methods by using them for low-resource dependency parsing, and discover that the joint training approach substantially outperforms the retrofitting approach. We further apply multilingual CWRs produced by the joint training approach to diverse languages, and show that it is still effective in transfer between distant languages, though we find that phylogenetically related source languages are generally more helpful.

We hypothesize that joint polyglot training is more successful than retrofitting because it induces a degree of lexical correspondence between languages that the linear transformation used in retrofitting methods cannot capture. To test this hypothesis, we design a *decontextual probe*. We *decontextualize* CWRs into non-contextual word vectors that retain much of CWRs' task-performance benefit, and evaluate the crosslingual transferability of language models via word translation. In our decontextualization framework, we use a single LSTM cell without recurrence to obtain a context-independent vector, thereby allowing for a direct probe into the LSTM networks in-

[*] Equal contribution. Random order.

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 304–315
Hong Kong, China, November 3-4, 2019.

dependent of a particular corpus. We show that decontextualized vectors from the joint training approach yield representations that score higher on a word translation task than the retrofitting approach or word type vectors such as fastText (Bojanowski et al., 2017). This finding provides evidence that polyglot language models encode crosslingual similarity, specifically crosslingual lexical correspondence, that a linear alignment between monolingual language models does not.

2 Models

We examine crosslingual solutions to low-resource dependency parsing, which make crucial use of multilingual CWRs. All models are implemented in AllenNLP, version 0.7.2 (Gardner et al., 2018) and the hyperparameters and training details are given in the appendix.

2.1 Multilingual CWRs

Prior methods to produce multilingual contextual word representations (CWRs) can be categorized into two major classes, which we call *joint training* and *retrofitting*.[1] The joint training approach trains a single polgylot language model (LM) on a mixture of texts in multiple languages (Mulcaire et al., 2019; Lample and Conneau, 2019; Devlin et al., 2019),[2] while the retrofitting approach trains separate LMs on each language and aligns the learned representations later (Schuster et al., 2019; Aldarmaki and Diab, 2019). We compare example approaches from these two classes using the same LM training data, and discover that the joint training approach generally yields better performance in low-resource dependency parsing, even without crosslingual supervision.

Retrofitting Approach Following Schuster et al. (2019), we first train a bidirectional LM with two-layer LSTMs on top of character CNNs for each language (ELMo, Peters et al., 2018), and then align the monolingual LMs across languages. Denote the hidden state in the jth layer for word i in context c by $\mathbf{h}_{i,c}^{(j)}$. We use a trainable weighted average of the three layers (character-CNN and two LSTM layers) to compute the contextual representation $\mathbf{e}_{i,c}$ for the word: $\mathbf{e}_{i,c} = \sum_{j=0}^{2} \lambda_j \mathbf{h}_{i,c}^{(j)}$ (Peters et al., 2018).[3] In the first step, we compute an "anchor" $\mathbf{h}_i^{(j)}$ for each word by averaging $\mathbf{h}_{i,c}^{(j)}$ over all occurrences in an LM corpus. We then apply a standard dictionary-based technique[4] to create multilingual word embeddings (Mikolov et al., 2013; Conneau et al., 2018). In particular, suppose that we have a word-translation dictionary from source language s to target language t. Let $\mathbf{H}_s^{(j)}, \mathbf{H}_t^{(j)}$ be matrices whose columns are the anchors in the jth layer for the source and corresponding target words in the dictionary. For each layer j, find the linear transformation $\mathbf{W}^{*(j)}$ such that

$$\mathbf{W}^{*(j)} = \operatorname*{argmin}_{\mathbf{W}} ||\mathbf{W}\mathbf{H}_s^{(j)} - \mathbf{H}_t^{(j)}||_F$$

The linear transformations are then used to map the LM hidden states for the source language to the target LM space. Specifically, contextual representations for the source and target languages are computed by $\sum_{j=0}^{2} \lambda_j \mathbf{W}^{*(j)} \mathbf{h}_{i,c}^{(j)}$ and $\sum_{j=0}^{2} \lambda_j \mathbf{h}_{i,c}^{(j)}$ respectively. We use publicly available dictionaries from Conneau et al. (2018)[5] and align all languages to the English LM space, again following Schuster et al. (2019).

Joint Training Approach Another approach to multilingual CWRs is to train a single LM on multiple languages (Tsvetkov et al., 2016; Ragni et al., 2016; Östling and Tiedemann, 2017). We train a single bidirectional LM with charater CNNs and two-layer LSTMs on multiple languages (Rosita, Mulcaire et al., 2019). We then use the polyglot LM to provide contextual representations. Similarly to the retrofitting approach, we represent word i in context c as a trainable weighted average of the hidden states in the trained polyglot LM: $\sum_{j=0}^{2} \lambda_j \mathbf{h}_{i,c}^{(j)}$. In contrast to retrofitting, crosslinguality is learned implicitly by sharing all network parameters during LM training; no crosslingual dictionaries are used.

[1]This term was originally used by Faruqui et al. (2015) to describe updates to word vectors, after estimating them from corpora, using semantic lexicons. We generalize it to capture the notion of a separate update to fit something other than the original data, applied after conventional training.

[2]Multilingual BERT is documented in `https://github.com/google-research/bert/blob/master/multilingual.md`.

[3]Schuster et al. (2019) only used the first LSTM layer, but we found a performance benefit from using all layers in preliminary results.

[4]Conneau et al. (2018) developed an unsupervised alignment technique that does not require a dictionary. We found that their unsupervised alignment yielded substantially degraded performance in downstream parsing in line with the findings of Schuster et al. (2019).

[5]`https://github.com/facebookresearch/MUSE#ground-truth-bilingual-dictionaries`

Refinement after Joint Training It is possible to combine the two approaches above; the alignment procedure used in the retrofitting approach can serve as a refinement step on top of an already-polyglot language model. We will see only a limited gain in parsing performance from this refinement in our experiments, suggesting that polyglot LMs are already producing high-quality multilingual CWRs even without crosslingual dictionary supervision.

FastText Baseline We also compare the multilingual CWRs to a subword-based, non-contextual word embedding baseline. We train 300-dimensional word vectors on the same LM data using the fastText method (Bojanowski et al., 2017), and use the same bilingual dictionaries to align them (Conneau et al., 2018).

2.2 Dependency Parsers

We train polyglot parsers for multiple languages (Ammar et al., 2016) on top of multilingual CWRs. All parser parameters are shared between the source and target languages. Ammar et al. (2016) suggest that sharing parameters between languages can alleviate the low-resource problem in syntactic parsing, but their experiments are limited to (relatively similar) European languages. Mulcaire et al. (2019) also include experiments with dependency parsing using polyglot contextual representations between two language pairs (English/Chinese and English/Arabic), but focus on high-resource tasks. Here we explore a wider range of languages, and analyze the particular efficacy of a crosslingual approach to dependency parsing in a low-resource setting.

We use a strong graph-based dependency parser with BiLSTM and biaffine attention (Dozat and Manning, 2017), which is also used in related work (Schuster et al., 2019; Mulcaire et al., 2019). Crucially, our parser only takes as input word representations. Universal parts of speech have been shown useful for low-resource dependency parsing (Duong et al., 2015; Ammar et al., 2016; Ahmad et al., 2019), but many realistic low-resource scenarios lack reliable part-of-speech taggers; here, we do not use parts of speech as input, and thus avoid the error-prone part-of-speech tagging pipeline. For the fastText baseline, word embeddings are not updated during training, to preserve crosslingual alignment (Ammar et al., 2016).

Lang	Code	Genus	WALS 81A	Size
English	ENG	Germanic	SVO	–
Arabic	ARA	Semitic	VSO/SVO	5241
Hebrew	HEB	Semitic	SVO	
Croatian	HRV	Slavic	SVO	6983
Russian	RUS	Slavic	SVO	
Dutch	NLD	Germanic	SOV/SVO	12269
German	DEU	Germanic	SOV/SVO	
Spanish	SPA	Romance	SVO	12543
Italian	ITA	Romance	SVO	
Chinese	CMN	Chinese	SVO	3997
Japanese	JPN	Japanese	SOV	
Hungarian	HUN	Ugric	SOV/SVO	910
Finnish	FIN	Finnic	SVO	12217
Vietnamese	VIE	Viet-Muong	SVO	1400
Uyghur	UIG	Turkic	SOV	1656
Kazakh	KAZ	Turkic	SOV	31
Turkish	TUR	Turkic	SOV	3685

Table 1: List of the languages used in our UD v2.2 experiments. Each shaded/unshaded section corresponds to a pair of "related" languages. WALS 81A denotes Feature 81A in WALS, Order of Subject, Object, and Verb (Dryer and Haspelmath, 2013). "Size" represents the downsampled size in # of sentences used for source treebanks. The four languages in bold face are truly low resource languages (< 2000 trees).

3 Experiments

We first conduct a set of experiments to assess the efficacy of multilingual CWRs for low-resource dependency parsing.

3.1 Zero-Target Dependency Parsing

Following prior work on low-resource dependency parsing and crosslingual transfer (Zhang and Barzilay, 2015; Guo et al., 2015; Ammar et al., 2016; Schuster et al., 2019), we conduct multi-source experiments on six languages (German, Spanish, French, Italian, Portuguese, and Swedish) from Google universal dependency treebank version 2.0 (McDonald et al., 2013).[6] We train language models on the six languages and English to produce multilingual CWRs. For each tested language, we train a polyglot parser with the multilingual CWRs on the five other languages and English, and apply the parser to the test data for the target language. Importantly, the parsing annotation scheme is shared among the seven languages. Our results will show that the joint training approach for CWRs substantially outperforms the retrofitting approach.

[6] http://github.com/ryanmcd/uni-dep-tb

3.2 Diverse Low-Resource Parsing

The previous experiment compares the joint training and retrofitting approaches in low-resource dependency parsing only for relatively similar languages. In order to study the effectiveness more extensively, we apply it to a more typologically diverse set of languages. We use five pairs of languages for "low-resource simulations," in which we reduce the size of a large treebank, and four languages for "true low-resource experiments," where only small UD treebanks are available, allowing us to compare to other work in the low-resource condition (Table 1). Following de Lhoneux et al. (2018), we selected these language pairs to represent linguistic diversity. For each target language, we produce multilingual CWRs by training a polyglot language model with its related language (e.g., Arabic and Hebrew) as well as English (e.g., Arabic and English). We then train a polyglot dependency parser on each language pair and assess the crosslingual transfer in terms of target parsing accuracy.

Each pair of related languages shares features like word order, morphology, or script. For example, Arabic and Hebrew are similar in their rich transfixing morphology (de Lhoneux et al., 2018), and Dutch and German share most of their word order features. We chose Chinese and Japanese as an example of a language pair which does *not* share a language family but does share characters.

We chose Hungarian, Vietnamese, Uyghur, and Kazakh as true low-resource target languages because they had comparatively small amounts of annotated text in the UD corpus (Vietnamese: 1,400 sentences, 20,285 tokens; Hungarian: 910 sentences, 20,166 tokens; Uyghur: 1,656 sentences, 19,262 tokens; Kazakh: 31 sentences, 529 tokens;), yet had convenient sources of text for LM pretraining (Zeman et al., 2018).[7] Other small treebanks exist, but in most cases another larger treebank exists for the same language, making domain adaptation a more likely option than crosslingual transfer. Also, recent work (Che et al., 2018) using contextual embeddings was top-ranked for most of these languages in the CoNLL 2018 shared task on UD parsing (Zeman et al., 2018).[8]

We use the same Universal Dependencies (UD) treebanks (Nivre et al., 2018) and train/development/test splits as the CoNLL 2018 shared task (Zeman et al., 2018).[9] The annotation scheme is again shared across languages, which facilitates crosslingual transfer. For each triple of two related languages and English, we downsample training and development data to match the language with the smallest treebank size. This allows for fairer comparisons because within each triple, the source language for any parser will have the same amount of training data. We further downsample sentences from the target train/development data to simulate low-resource scenarios. The ratio of training and development data is kept 5:1 throughout the simulations, and we denote the number of sentences in training data by $|D_\tau|$. For testing, we use the CoNLL 2018 script on the gold word segmentations. For the truly low-resource languages, we also present results with word segmentations from the system outputs of Che et al. (2018) (HUN, VIE, UIG) and Smith et al. (2018) (KAZ) for a direct comparison to those languages' best previously reported parsers.[10]

4 Results and Discussion

In this section we describe the results of the various parsing experiments.

4.1 Zero-Target Parsing

Table 2 shows results on zero-target dependency parsing. First, we see that all CWRs greatly improve upon the fastText baseline. The joint training approach (Rosita), which uses no dictionaries, consistently outperforms the dictionary-dependent retrofitting approach (ELMos+Alignment). As discussed in the previous section, we can apply the alignment method to refine the already-polyglot Rosita using dictionaries. However, we observe a relatively limited gain in overall performance (74.5 vs. 73.9 LAS points), suggesting that Rosita (polyglot language model) is already developing useful multilingual CWRs for parsing without crosslingual supervision. Note that the degraded overall performance of our ELMo+Alignment compared to Schuster et al. (2019)'s reported results (71.2 vs. 73.1) is likely

[7]The one exception is Uyghur where we only have 3M words in the raw LM data from Zeman et al. (2018).

[8]In Kazakh, Che et al. (2018) did not use CWRs due to the extremely small treebank size.

[9]See Appendix for a list of UD treebanks used.

[10]System outputs for all shared task systems are available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2885

Model	DEU	SPA	FRA	ITA	POR	SWE	AVG
Schuster et al. (2019) (retrofitting)	61.4	77.5	**77.0**	77.6	73.9	71.0	73.1
Schuster et al. (2019) (retrofitting, no dictionaries)	**61.7**	76.6	76.3	**77.1**	69.1	54.2	69.2
fastText + Alignment	45.2	68.5	62.8	58.9	61.1	50.4	57.8
ELMos + Alignment (retrofitting)	57.3	75.4	73.7	71.6	75.1	74.2	71.2
Rosita (joint training, no dictionaries)	58.0	**81.8**	75.6	74.8	**77.1**	76.2	73.9
Rosita + Refinement (joint training + retrofitting)	**61.7**	79.7	75.8	76.0	76.8	**76.7**	**74.5**

Table 2: Zero-target results in LAS. Results reported in prior work (above the line) use an unknown amount of LM training data; all models below the line are limited to approximately 50M words per language.

Model	DEU	SPA	FRA	ITA	POR	SWE	AVG
Zhang and Barzilay (2015)	54.1	68.3	68.8	69.4	72.5	62.5	65.9
Guo et al. (2016)	55.9	73.1	71.0	71.2	78.6	69.5	69.9
Ammar et al. (2016)	57.1	74.6	73.9	72.5	77.0	68.1	70.5
Schuster et al. (2019) (retrofitting)	**65.2**	80.0	**80.8**	**79.8**	82.7	75.4	77.3
Schuster et al. (2019) (retrofitting, no dictionaries)	64.1	77.8	79.8	79.7	79.1	69.6	75.0
Rosita (joint training, no dictionaries)	63.6	**83.4**	78.9	77.8	83.0	**79.6**	77.7
Rosita + Refinement (joint training + retrofitting)	64.8	82.1	78.7	78.8	**84.1**	79.1	**77.9**

Table 3: Zero-target results in LAS with gold UPOS.

due to the significantly reduced amount of LM data we used in all of our experiments (50M words per language, an order of magnitude reduction from the full Wikipedia dumps used in Schuster et al. (2019)). Schuster et al. (2019) (no dictionaries) is the same retrofitting approach as ELMos+Alignment except that the transformation matrices are learned in an unsupervised fashion without dictionaries (Conneau et al., 2018). The absence of a dictionary yields much worse performance (69.2 vs. 73.1) in contrast with the joint training approach of Rosita, which also does not use a dictionary (73.9).

We also present results using gold universal part of speech to compare to previous work in Table 3. We again see Rosita's effectiveness and a marginal benefit from refinement with dictionaries. It should also be noted that the reported results for French, Italian and German in Schuster et al. (2019) outperform all results from our controlled comparison; this may be due to the use of abundant LM training data. Nevertheless, joint training, with or without refinement, performs best on average in both gold and predicted POS settings.

4.2 Diverse Low-Resource Parsing

Low-Resource Simulations Figure 1 shows simulated low-resource results.[11] Of greatest interest are the significant improvements over monolingual parsers when adding English or related-language data. This improvement is consistent across languages and suggests that crosslingual transfer is a viable solution for a wide range of languages, even when (as in our case) language-specific tuning or annotated resources like parallel corpora or bilingual dictionaries are not available. See Figure 2 for a visualization of the differences in performance with varying training size. The polyglot advantage is minor when the target language treebank is large, but dramatic in the condition where the target language has only 100 sentences. The fastText approaches consistently underperform the language model approaches, but show the same pattern.

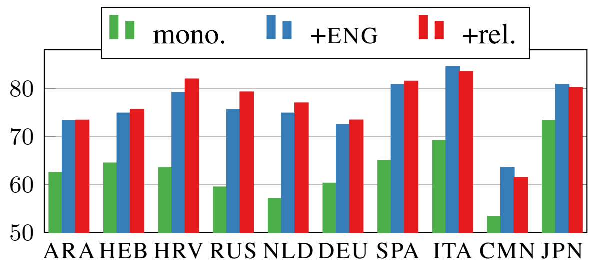

Figure 1: LAS for UD parsing results in a simulated low-resource setting where the size of the target language treebank ($|D_\tau|$) is set to 100 sentences.

In addition, related-language polyglot ("+rel.") outperforms English polyglot in most cases in the low-resource condition. The exceptions to this pattern are Italian (whose treebank is of a different genre from the Spanish one), and Japanese and Chinese, which differ significantly in morphology and word order. The CMN/JPN result suggests that such typological features influence the degree of crosslingual transfer more than orthographic

[11] A table with full details including different size simulations is provided in the appendix.

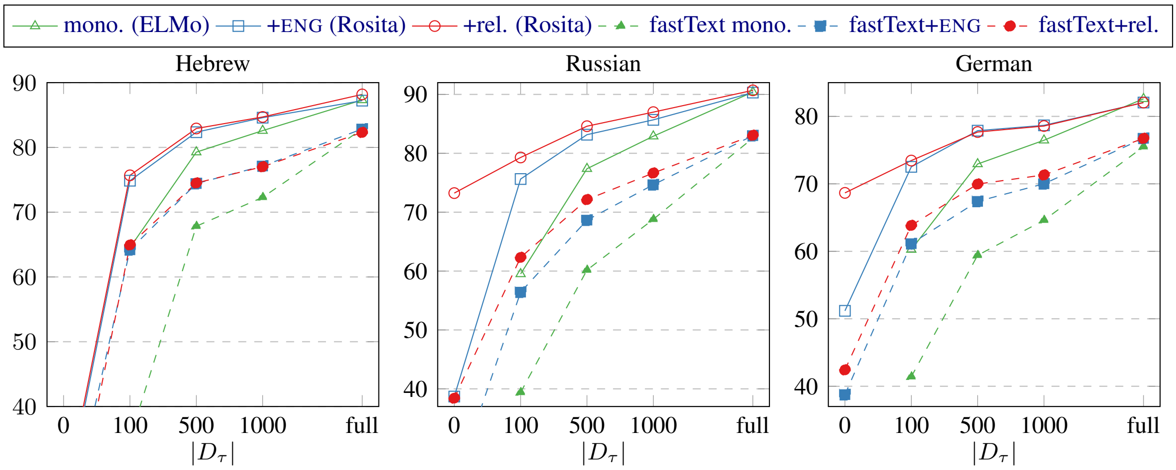

Figure 2: Plots of parsing performance vs. target language treebank size for several example languages. The size 0 target treebank point indicates a parser trained *only* on the source language treebank but with polyglot representations, allowing transfer to the target test treebank using no target language training trees. See Appendix for results with zero-target-treebank and intermediate size data ($|D_\tau| \in \{0, 100, 500, 1000\}$) for all languages.

properties like shared characters. This result in crosslingual transfer also mirrors the observation from prior work (Gerz et al., 2018) that typological features of the language are predictive of *monolingual* LM performance. The related-language improvement also vanishes in the full-data condition (Figure 2), implying that the importance of shared linguistic features can be overcome with sufficient annotated data. It is also noteworthy that variations in word order, such as the order of adjective and noun, do not affect performance: Italian, Arabic, and others use a noun-adjective order while English uses an adjective-noun order, but their +ENG and +rel. results are comparable.

The Croatian and Russian results are notable because of shared heritage but different scripts. Though Croatian uses the Latin alphabet and Russian uses Cyrillic, transfer between HRV+RUS is clearly more effective than HRV+ENG (82.00 vs. 79.21 LAS points when $|D_\tau| = 100$). This suggests that character-based LMs can implicitly learn to transliterate between related languages with different scripts, even without parallel supervision.

Truly Low Resource Languages Finally we present "true low-resource" experiments for four languages in which little UD data is available (see Section 3.2). Table 4 shows these results. Consistent with our simulations, our parsers on top of Rosita (multilingual CWRs from the joint training approach) substantially outperform the parsers with ELMos (monolingual CWRs) in all languages, and establish a new state of the art

Model	gold	pred.
Hungarian (HUN)		
Che et al. (2018) (HUN, ensemble)	–	82.66
Che et al. (2018) (HUN)	–	80.96
ELMo (HUN)	81.89	81.54
Rosita (HUN +ENG)	85.34	84.89
Rosita (HUN +FIN)	**85.40**	**84.96**
Vietnamese (VIE)		
Che et al. (2018) (VIE, ensemble)	–	55.22
ELMo (VIE)	62.67	55.72
Rosita (VIE +ENG)	**63.07**	**56.42**
Uyghur (UIG)		
Che et al. (2018) (UIG, ensemble)	–	**67.05**
Che et al. (2018) (UIG)	–	66.20
ELMo (UIG)	66.64	63.98
Rosita (UIG +ENG)	67.85	65.55
Rosita (UIG +TUR)	**68.08**	65.73
Kazakh (KAZ)		
Rosa and Mareček (2018) (KAZ +TUR)	–	26.31
Smith et al. (2018) (KAZ +TUR)	–	31.93
Schuster et al. (2019) (KAZ +TUR)	–	36.98
Rosita (KAZ +ENG)	48.02	46.03
Rosita (KAZ +TUR)	**53.98**	**51.96**

Table 4: LAS (F_1) comparison for truly low-resource languages. The gold and pred. columns show results under gold segmentation and predicted segmentation. The languages in the parentheses indicate the languages used in parser training.

in Hungarian, Vietnamese, and Kazakh. Consistent with our simulations, we see that training parsers with the target's related language is more effective than with the more distant language, English. It is particularly noteworthy that the Rosita models, which do not use a parallel corpus or dictionary, dramatically improve over the best previously reported result from Schuster et al. (2019) when either the related language of Turkish (51.96 vs. 36.98) or even the more distant language of English (46.03 v.s. 36.98) is used. Schuster et al. (2019) aligned the monolingual ELMos for Kazakh and Turkish using the KAZ-TUR dictionary that Rosa and Mareček (2018) derived from parallel text. This result further corroborates our finding that the joint training approach to multilingual CWRs is more effective than retrofitting monolingual LMs.

4.3 Comparison to Multilingual BERT Embeddings

We also evaluate the diverse low-resource language pairs using pretrained multilingual BERT (Devlin et al., 2019) as text embeddings (Figure 3). Here, the same language model (multilingual cased BERT,[12] covering 104 languages) is used for all parsers, with the only variation being in the training treebanks provided to each parser. Parsers are trained using the same hyperparameters and data as in Section 3.2.[13]

There are two critical differences from our previous experiments: multilingual BERT is trained on much larger amounts of Wikipedia data compared to other LMs used in this work, and the WordPiece vocabulary (Wu et al., 2016) used in the cased multilingual BERT model has been shown to have a distribution skewed toward Latin alphabets (Ács, 2019). These results are thus not directly comparable to those in Figure 1; nevertheless, it is interesting to see that the results obtained with ELMo-like LMs are comparable to and in some cases better than results using a BERT model trained on over a hundred languages. Our results broadly fit with those of Pires et al. (2019), who found that multilingual BERT was useful for zero-shot crosslingual syntactic transfer. In particular, we find nearly no performance benefit from cross-script transfer using BERT in a language pair (English-Japanese) for which they reported

[12]Available at https://github.com/google-research/bert/

[13]AllenNLP version 0.9.0 was used for these experiments.

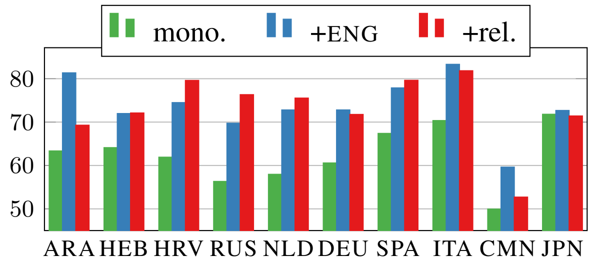

Figure 3: LAS for UD parsing results in a simulated low-resource setting (($|D_\tau| = 100$) using multilingual BERT embeddings in place of Rosita. Cf. Figure 1.

poor performance in zero-shot transfer, contrary to our results using Rosita (Section 4.2).

5 Decontextual Probe

We saw the success of the joint polyglot training for multilingual CWRs over the retrofitting approach in the previous section. We hypothesize that CWRs from joint training provide useful representations for parsers by inducing *nonlinear* similarity in the vector spaces of different languages that we cannot retrieve with a simple alignment of monolingual pretrained language models. In order to test this hypothesis, we conduct a *decontextual probe* comprised of two steps. The decontextualization step effectively distills CWRs into word type vectors, where each unique word is mapped to exactly one embedding regardless of the context. We then conduct linear transformation-based word translation (Mikolov et al., 2013) on the decontextualized vectors to quantify the degree of crosslingual similarity in the multilingual CWRs.

5.1 Decontextualization

Recall from Section 2 that we produce CWRs from bidirectional LMs with character CNNs and two-layer LSTMs. We propose a method to remove the dependence on context c for the two LSTM layers (the CNN layer is already context-independent by design). During LM training, the hidden states of each layer h_t are computed by the standard LSTM equations:

$$
\begin{aligned}
i_t &= \sigma\left(W_i x_t + +U_i h_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\
\tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\
o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh\left(c_t\right)
\end{aligned}
$$

Representations	UD	SRL	NER
GloVe	83.78	80.01	83.90
fastText	83.93	80.27	83.40
Decontextualization	86.88	81.41	87.72
ELMo	88.71	82.12	88.65

Table 5: Context independent vs. dependent performance in English. All embeddings are 512-dimensional and trained on the same English corpus of approximately 50M tokens for fair comparisons. We also concatenate 128-dimensional character LSTM representations with the word vectors in every configuration to ensure all models have character input. UD scores are LAS, and SRL and NER are F_1.

We produce contextless vectors from pretrained LMs by removing recursion in the computation (i.e. setting h_{t-1} and c_{t-1} to 0):

$$\begin{aligned} i_t &= \sigma\left(W_i x_t + b_i\right) \\ f_t &= \sigma\left(W_f x_t + b_f\right) \\ \tilde{c}_t &= \tanh\left(W_c x_t + b_c\right) \\ o_t &= \sigma\left(W_o x_t + b_o\right) \\ c_t &= i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned}$$

This method is fast to compute, as it does not require recurrent computation and only needs to see each word once. This way, each word is associated with a set of exactly three vectors from the three layers.

Performance of decontextualized vectors We perform a brief experiment to find what information is successfully retained by the decontextualized vectors, by using them as inputs to three tasks (in a monolingual English setting, for simplicity). For Universal Dependencies (UD) parsing, semantic role labeling (SRL), and named entity recognition (NER), we used the standard train/development/test splits from UD English EWT (Zeman et al., 2018) and Ontonotes (Pradhan et al., 2013). Following Mulcaire et al. (2019), we use strong existing neural models for each task: Dozat and Manning (2017) for UD parsing, He et al. (2017) for SRL, and Peters et al. (2017) for NER.

Table 5 compares the decontextualized vectors with the original CWRs (ELMo) and the conventional word type vectors, GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). In all three tasks, the decontextualized vectors substantially improve over fastText and GloVe vectors, and perform nearly on par with contextual

Vector	DEU	SPA	FRA	ITA	POR	SWE
fastText	31.6	54.8	56.7	50.2	55.5	43.9
ELMos						
Layer 0	19.7	41.5	41.1	36.9	44.6	27.5
Layer 1	24.4	46.4	47.6	44.2	48.3	36.3
Layer 2	19.9	40.5	41.9	38.1	42.5	30.9
Rosita						
Layer 0	37.9	**56.6**	**58.2**	57.5	**56.6**	50.6
Layer 1	**40.3**	56.3	57.2	**58.1**	56.5	**53.7**
Layer 2	38.8	51.1	52.7	53.6	50.7	50.8

Table 6: Crosslingual alignment results (precision at 1) from decontextual probe. Layers 0, 1, and 2 denote the character CNN, first LSTM, and second LSTM layers in the language models respectively.

ELMo. This suggests that while part of the advantage of CWRs is in the incorporation of context, they also benefit from rich context-*independent* representations present in deeper networks.

5.2 Word Translation Test

Given the decontextualized vectors from each layer of the bidirectional language models, we can measure the crosslingual lexical correspondence in the multilingual CWRs by performing word translation. Concretely, suppose that we have training and evaluation word translation pairs from the source to the target language. Using the same word alignment objective discussed as in Section 2.1, we find a linear transform by aligning the decontextualized vectors for the training source-target word pairs. Then, we apply this linear transform to the decontextualized vector for each source word in the evaluation pairs. The closest target vector is found using the cross-domain similarity local scaling (CSLS) measure (Conneau et al., 2018), which is designed to remedy the hubness problem (where a few "hub" points are nearest neighbors to many other points each) in word translation by normalizing the cosine similarity according to the degree of hubness.

We again take the dictionaries from Conneau et al. (2018) with the given train/test split, and always use English as the target language. For each language, we take all words that appear three times or more in our LM training data and compute decontextualized vectors for them. Word translation is evaluated by choosing the closest vector among the English decontextualized vectors.

5.3 Results

We present word translation results from our decontextual probe in Table 6. We see that the first

LSTM layer generally achieves the best crosslingual alignment both in ELMos and Rosita. This finding mirrors recent studies on layerwise transferability; representations from the first LSTM layer in a language model are most transferable across a range of tasks (Liu et al., 2019). Our decontextual probe demonstrates that the first LSTM layer learns the most generalizable representations not only across tasks but also across languages. In all six languages, Rosita (joint LM training approach) outperforms ELMos (retrofitting approach) and the fastText vectors. This shows that for the polyglot (jointly trained) LMs, there is a preexisting similarity between languages' vector spaces beyond what a linear transform provides. The resulting language-agnostic representations lead to polyglot training's success in low-resource dependency parsing.

6 Further Related Work

In addition to the work mentioned above, much previous work has proposed techniques to transfer knowledge from a high-resource to a low-resource language for dependency parsing. Many of these methods use an essentially (either lexicalized or delexicalized) joint polyglot training setup (e.g., McDonald et al., 2011; Cohen et al., 2011; Duong et al., 2015; Guo et al., 2016; Vilares et al., 2016; Falenska and Çetinoğlu, 2017 as well as many of the CoNLL 2017/2018 shared task participants: Lim and Poibeau (2017); Vania et al. (2017); de Lhoneux et al. (2017); Che et al. (2018); Wan et al. (2018); Smith et al. (2018); Lim et al. (2018)). Some use typological information to facilitate crosslingual transfer (e.g., Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015; Wang and Eisner, 2016; Rasooli and Collins, 2017; Ammar et al., 2016). Others use bitext (Zeman et al., 2018), manually-specified rules (Naseem et al., 2012), or surface statistics from gold universal part of speech (Wang and Eisner, 2018a,b) to map the source to target. The methods examined in this work to produce multilingual CWRs do not rely on such external information about the languages, and instead use relatively abundant LM data to learn crosslinguality that abstracts away from typological divergence.

Recent work has developed several probing methods for (monolingual) contextual representations (Liu et al., 2019; Hewitt and Manning, 2019; Tenney et al., 2019). Wada and Iwata (2018) showed that the (contextless) input and output word vectors in a polyglot word-based language model manifest a certain level of lexical correspondence between languages. Our decontextual probe demonstrated that the *internal* layers of polyglot language models capture crosslinguality and produce useful multilingual CWRs for downstream low-resource dependency parsing.

7 Conclusion

We assessed recent approaches to multilingual contextual word representations, and compared them in the context of low-resource dependency parsing. Our parsing results illustrate that a joint training approach for polyglot language models outperforms a retrofitting approach of aligning monolingual language models. Our decontextual probe showed that jointly trained LMs learn a better crosslingual lexical correspondence than the one produced by aligning monolingual language models or word type vectors. Our results provide a strong basis for multilingual representation learning and for further study of crosslingual transfer in a low-resource setting beyond dependency parsing.

Acknowledgments

The authors thank Nikolaos Pappas and Tal Schuster as well as the anonymous reviewers for their helpful feedback. This research was funded in part by NSF grant IIS-1562364, a Google research award to NAS, the Funai Overseas Scholarship to JK, and the NVIDIA Corporation through the donation of a GeForce GPU.

References

Judit Ács. 2019. Exploring BERT's vocabulary.

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proc. of NAACL-HLT*.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proc. of NAACL-HLT*.

Waleed Ammar. 2016. *Towards a Universal Analyzer of Natural Languages*. Ph.D. thesis, Carnegie Mellon University.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *TACL*, 4.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5.

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 *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. arXiv:1312.3005.

Shay B Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proc. of EMNLP*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proc. of ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*.

Timothy Dozat and Christopher Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proc. of ICLR*.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proc. of ACL-IJCNLP*.

Agnieszka Falenska and Özlem Çetinoğlu. 2017. Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proc. of IWPT*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proc. of NAACL-HLT*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proc. of NLP-OSS*.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proc. of EMNLP*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proc. of ACL-IJCNLP*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Proc. of AAAI*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proc. of ACL*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proc. of NAACL-HLT*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. ADAM: A Method for Stochastic Optimization. In *Proc. of ICLR*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proc. of NeurIPS*.

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proc. of EMNLP*.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017. From raw text to universal dependencies - look, no tags! In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

KyungTae Lim, Cheoneum Park, Changki Lee, and Thierry Poibeau. 2018. SEx BiST: A multi-source trainable parser with deep contextualized lexical representations. In *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

KyungTae Lim and Thierry Poibeau. 2017. A system for multilingual dependency parsing based on bidirectional LSTM feature representations. In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proc. of NAACL-HLT*.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proc. of ACL*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proc. of ACL*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proc. of EMNLP*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. arXiv:1309.4168.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proc. of NAACL-HLT*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proc. of ACL*.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proc. of EACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of EMNLP*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proc. of ACL*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proc. of ACL*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proc. of CoNLL*.

Anton Ragni, Edgar Dakin, Xie Chen, Mark J. F. Gales, and Kate Knill. 2016. Multi-language neural network language models. In *Proc. of INTERSPEECH*.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *TACL*, 5.

Rudolf Rosa and David Mareček. 2018. CUNI x-ling: Parsing under-resourced languages in CoNLL 2018 UD shared task. In *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proc. of NAACL-HLT*.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proc. of NAACL-HLT*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proc. of ACL*.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proc. of NAACL-HLT*.

Clara Vania, Xingxing Zhang, and Adam Lopez. 2017. UParse: the Edinburgh system for the CoNLL 2017 UD shared task. In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. In *Proc. of ACL*.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. arXiv:1809.02306.

Hui Wan, Tahira Naseem, Young-Suk Lee, Vittorio Castelli, and Miguel Ballesteros. 2018. IBM research at the CoNLL 2018 shared task on multilingual parsing. In *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Dingquan Wang and Jason Eisner. 2016. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *TACL*, 4.

Dingquan Wang and Jason Eisner. 2018a. Surface statistics of an unknown language indicate how to parse it. *TACL*, 6.

Dingquan Wang and Jason Eisner. 2018b. Synthetic data made to order: The case of parsing. In *Proc. of EMNLP*.

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:1609.08144.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. arxiv:1212.5701.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proc. of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proc. of EMNLP*.