Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank

Ethan C. Chau[†◇] **Lucy H. Lin**[†] **Noah A. Smith**[†⋆]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Department of Linguistics, University of Washington
[⋆]Allen Institute for Artificial Intelligence
{echau18,lucylin,nasmith}@cs.washington.edu

Abstract

Pretrained multilingual contextual representations have shown great success, but due to the limits of their pretraining data, their benefits do not apply equally to all language varieties. This presents a challenge for language varieties unfamiliar to these models, whose labeled *and unlabeled* data is too limited to train a monolingual model effectively. We propose the use of additional language-specific pretraining and vocabulary augmentation to adapt multilingual models to low-resource settings. Using dependency parsing of four diverse low-resource language varieties as a case study, we show that these methods significantly improve performance over baselines, especially in the lowest-resource cases, and demonstrate the importance of the relationship between such models' pretraining data and target language varieties.

1 Introduction

Contextual word representations (CWRs) from pretrained language models have improved many NLP systems. Such language models include BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), which are conventionally "pretrained" on large unlabeled datasets before their internal representations are "finetuned" during supervised training on downstream tasks like parsing. However, many language varieties[1] lack large annotated and even unannotated datasets, raising questions about the broad applicability of such data-hungry methods.

One exciting way to compensate for the lack of unlabeled data in low-resource language varieties is to finetune a large, *multilingual* language model that has been pretrained on the union of many languages' data (Devlin et al., 2019; Lample and Conneau, 2019). This enables the model to transfer some of what it learns from high-resource languages to low-resource ones, demonstrating benefits over monolingual methods in some cases (Conneau et al., 2020a; Tsai et al., 2019), though not always (Agerri et al., 2020; Rönnqvist et al., 2019).

Specifically, multilingual models face the transfer-dilution tradeoff (Conneau et al., 2020a): increasing the number of languages during pretraining improves positive crosslingual transfer but decreases the model capacity allocated to each language. Furthermore, such models are only pretrained on a finite amount of data and may lack exposure to specialized domains of certain languages or even entire low-resource language varieties. The result is a challenge for these language varieties, which must rely on positive transfer from a sufficient number of similar high-resource languages. Indeed, Wu and Dredze (2020) find that multilingual models often underperform monolingual baselines for such languages and question their off-the-shelf viability.

We take inspiration from previous work on domain adaptation, where general-purpose monolingual models have been effectively adapted to specialized domains through additional pretraining on domain-specific corpora (Gururangan et al., 2020). We hypothesize that we can improve the performance of multilingual models on low-resource language varieties analogously, through additional pretraining on *language*-specific corpora.

However, additional pretraining on more data in the target language does not ensure its full representation in the model's vocabulary, which is constructed to maximally represent the model's original pretraining data (Sennrich et al., 2016; Wu et al., 2016). Artetxe et al. (2020) find that target languages' representation in the vocabulary affects these models' transferability, suggesting that language varieties on the fringes of the vocabulary

[1] Sociolinguists define "language varieties" broadly to encompass any distinct form of a language. In addition to standard varieties (conventionally referred to as "languages"), this includes dialects, registers, and styles (Trudgill, 2003).

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1324–1334
November 16 - 20, 2020.

may not be sufficiently well-modeled. Can we incorporate vocabulary from the target language into multilingual models' existing alignment?

We introduce the use of additional language-specific pretraining for multilingual CWRs in a low-resource setting, *before* use in a downstream task; to better model language-specific tokens, we also augment the existing vocabulary with frequent tokens from the low-resource language (§2). Our experiments consider dependency parsing in four typologically diverse low-resource language varieties with different degrees of relatedness to a multilingual model's pretraining data (§3). Our results show that these methods consistently improve performance on each target variety, especially in the lowest-resource cases (§4). In doing so, we demonstrate the importance of accounting for the relationship between a multilingual model's pretraining data and the target language variety.

Because the pretraining-finetuning paradigm is now ubiquitous, many experimental findings for one task can now inform work on other tasks. Thus, our findings on dependency parsing—whose annotated datasets cover many more low-resource language varieties than those of other NLP tasks—are expected to interest researchers and practitioners facing low-resource situations for other tasks. To this end, we make our code, data, and hyperparameters publicly available.[2]

2 Overview

We are chiefly concerned with the adaptation of pretrained multilingual models to a target language by optimally using available data. As a case study, we use the multilingual cased BERT model (MBERT) of Devlin et al. (2019), a transformer-based (Vaswani et al., 2017) language model which has produced strong CWRs for many languages (Kondratyuk and Straka, 2019, *inter alia*). MBERT is pretrained on the 104 languages with the most Wikipedia data and encodes input tokens using a fixed wordpiece vocabulary (Wu et al., 2016) learned from this data. Low-resource languages are slightly oversampled in its pretraining data, but high resource languages are still more prevalent, resulting in a language imbalance.[3]

We observe that two types of target language varieties may be disadvantaged by this training scheme: the lowest-resource languages in MBERT's pretraining data (which we call Type 1); and unseen low-resource languages (Type 2). Although Type 1 languages are oversampled during training, they are still overshadowed by high-resource languages. Type 2 languages must rely purely on crosslingual vocabulary overlap. In both cases, the wordpieces that encode the input tokens in these languages may not fully capture the senses in which they are used, or they may be completely unseen.[4] However, other low-resource varieties with more representation in MBERT's pretraining data (Type 0) may not be as disadvantaged. Optimally using MBERT in low-resource settings thus requires accounting for limitations with respect to a target language variety.

2.1 Methods

We evaluate three methods of adapting MBERT to better model target language varieties.

Language-Adaptive Pretraining (LAPT) Under the assumption that language varieties function analagously to domains for MBERT, we adapt the *domain-adaptive pretraining* method of Gururangan et al. (2020) to a multilingual setting. With *language-adaptive pretraining*, MBERT is pretrained for additional epochs on monolingual data in the target language variety to improve the alignment of the wordpiece embeddings.

Vocabulary Augmentation (VA) To better model unseen or language-specific wordpieces, we explore performing LAPT after augmenting MBERT's vocabulary from a target language variety. We train a new wordpiece vocabulary on monolingual data in the target language, tokenize the monolingual data with the new vocabulary, and augment MBERT's vocabulary with the 99 most common wordpieces[5] in the new vocabulary that replaced the "unknown" wordpiece token. Full details of this process are given in the Appendix.

Tiered Vocabulary Augmentation (TVA) We consider a variant of VA with a larger learning rate

[2] https://github.com/ethch18/parsing-mbert

[3] Sampling is done based on an exponentially smoothed distribution of the amount of data in each language, which slightly increases the representation of low-resource languages. See https://github.com/google-research/bert/blob/master/multilingual.md for more details.

[4] Wordpiece tokenization is done greedily based on a fixed vocabulary. The model returns a special "unknown" token for unseen characters and other subword units that cannot be represented by the vocabulary.

[5] MBERT's fixed-size vocabulary contains 99 tokens designated as "unused," whose representations were not updated during initial pretraining and can be repurposed for vocabulary augmentation without modifying the pretrained model.

Language	Type	# Sentences	# Tokens	WP/Token	UNK Tokens
GA	1	199k	3.6M	2.10	12807
MT	2	62k	1M	2.95	49791
SING	0	80k	1.2M	1.24	3
VI	0	255k	5.6M	1.33	6955

Table 1: Unlabeled dataset statistics: number of sentences, number of tokens, average wordpieces per token, and tokens containing an unknown wordpiece under original MBERT vocabulary.

for the embeddings of the 99 new wordpieces than for the other parameters. We expect this method to learn the embeddings more thoroughly without overfitting the model's remaining parameters. Learning rate details are given in the Appendix.

2.2 Evaluation

We perform evaluation on dependency parsing. Following Kondratyuk and Straka (2019), we take a weighted sum of the activations at each MBERT layer as the CWR for each token. We then pass the representations into the graph-based dependency parser of Dozat and Manning (2017). This parser, which is also used in related work (Kondratyuk and Straka, 2019; Mulcaire et al., 2019a; Schuster et al., 2019), uses a biaffine attention mechanism between word representations to score a parse tree.

3 Experiments

We consider two variants of each MBERT method: one in which the pretrained CWRs are frozen; and one where they are further finetuned during parser training (FT). Following prior work involving these two variants (Beltagy et al., 2019), FT variants perform biaffine attention directly on the outputs of MBERT instead of first passing them through a BiLSTM, as in Dozat and Manning (2017).

We perform additional pretraining for up to 20 epochs, selecting our final models based on average validation LAS downstream. Full training details are given in the Appendix. We report average scores and standard errors based on five random initializations. Code and data are publicly available (see footnote 2).

3.1 Languages and Datasets

We perform experiments on four typologically diverse low-resource languages: Irish (GA), Maltese (MT), Vietnamese (VI), and Singlish (Singapore Colloquial English; SING). Singlish is an English-based creole spoken in Singapore, which incorporates lexical and syntactic borrowings from other languages spoken in Singapore: Chinese, Malay, and Tamil. Wang et al. (2017) provide an extended motivation for evaluating on Singlish.

These language varieties are examplars of the three types discussed in §2. MBERT is trained on the 104 largest Wikipedias, which includes Irish and Vietnamese but *excludes* Maltese and Singlish. However, the Irish Wikipedia is several orders of magnitude smaller than the full Vietnamese one. So, we view Irish and Maltese as Type 1 and Type 2 language varieties, respectively. Though Singlish lacks its own Wikipedia and is likely not included in MBERT's pretraining data *per se*, its component languages (English, Chinese, Malay, and Tamil) are all well-represented in the data. We thus consider it to be a Type 0 variety along with Vietnamese.

Unlabeled Datasets Additional pretraining for Irish, Maltese, and Vietnamese uses unlabeled articles from Wikipedia. To simulate a truly low-resource setting for Vietnamese, we use a random sample of 5% of the articles. Singlish data is crawled from the *SG Talk Forum*[6] online forum and provided by Wang et al. (2017). To ensure robust evaluation, we remove all sentences that appear in the labeled validation and test sets from the unlabeled data. Full details are provided in the Appendix.

Tab. 1 gives the average number of wordpieces per token and the number of tokens with unknown wordpieces in each of the unlabeled datasets, computed based on the original MBERT vocabulary. While the high number of wordpieces per token for Irish and Maltese may be due in part to morphological richness, it also suggests that these languages stand to benefit more from improved alignment of the wordpieces' embeddings. Furthermore, the higher rates of unknown wordpieces leave room for enhanced performance with an improved vocabulary.

[6] https://sgTalk.com

Representations	Irish (GA)	Maltese (MT)	Singlish (SING)	Vietnamese (VI)
	Type 1	Type 2	Type 0	Type 0
FASTT	65.36 ± 1.33	68.23 ± 0.61	66.42 ± 0.92	53.37 ± 0.95
ELMO	68.25 ± 0.37	74.33 ± 0.53	68.63 ± 1.04	56.91 ± 0.41
MBERT	68.19 ± 0.43	67.06 ± 0.61	74.01 ± 0.39	62.96 ± 0.41
LAPT	**73.03 ± 0.25**	78.51 ± 0.41	**76.48 ± 0.63**	**64.67 ± 0.22**
VA	72.68 ± 0.47	**79.88 ± 0.55**	**76.71 ± 0.70**	64.28 ± 0.44
TVA	**73.11 ± 0.37**	79.32 ± 0.45	**76.92 ± 0.77**	**64.46 ± 0.44**
MBERT + FT	72.67 ± 0.22	76.74 ± 0.35	78.24 ± 0.52	66.13 ± 0.38
LAPT + FT	75.45 ± 0.28	82.77 ± 0.24	79.30 ± 0.57	**67.50 ± 0.25**
VA + FT	**76.17 ± 0.08**	**83.53 ± 0.21**	**79.89 ± 0.46**	67.28 ± 0.38
TVA + FT	**76.23 ± 0.22**	83.16 ± 0.25	**80.09 ± 0.34**	**67.82 ± 0.27**

Table 2: Results (LAS) on downstream UD parsing, with standard deviations from five random initializations. **Bolded** results are within one standard deviation of the maximum for each category (frozen/FT).

Labeled Datasets Parsers for Irish, Maltese, and Vietnamese are trained on the corresponding treebanks and train/test splits from Universal Dependencies 2.5 (Zeman et al., 2019): IDT, MUDT, and VTB, respectively. For Singlish, we use the extended treebank component of Wang et al. (2019), which we randomly partition into train (80%), valid. (10%), and test sets (10%).[7] We use the provided gold word segmentation but no POS tag features.

3.2 Baselines

For each language, we evaluate the performance of MBERT in frozen and FT variants, without any adaptations. We additionally benchmark each method against strong prior work that represents conventional approaches for representing low-resource languages: static fastText embeddings (FASTT; Bojanowski et al., 2017), which can be learned effectively even on small datasets; and monolingual ELMo models (ELMO; Peters et al., 2018), a monolingual contextual approach. We choose ELMo over training a new BERT model because the high computational and data requirements of the latter make it unviable in a low-resource setting. Both baselines are trained on our unlabeled datasets.

4 Results and Discussion

Tab. 2 shows the performance of each of the method variants on the four Universal Dependencies datasets, with standard deviations from five different initializations. Our experiments demonstrate that additional language-specific pretraining results in more effective representations. LAPT consistently outperforms baselines, especially for Irish and Maltese, where overlap with the original pretraining data is low and frozen MBERT underperforms ELMO. This suggests that the insights of Gururangan et al. (2020) on additional pretraining for domain adaptation are also applicable to transferring multilingual models to low-resource languages, even without much additional data.

LAPT with our vocabulary augmentation methods yield small but significant improvements over LAPT alone, especially for FT configurations and Type 1/2 languages. This demonstrates that accurate vocabulary modeling is important for improving representations in the target language, and that VA and TVA are effective methods for doing so while maintaining overall alignment. For Maltese, VA's stronger performance compared to TVA can be explained by the overall lack of unlabeled data: one would expect TVA to overfit more quickly on a very small dataset.

Furthermore, the relative error reductions between unadapted MBERT and each of our methods correlates with each language variety's relationship to MBERT pretraining data. Maltese (Type 2) improves by up to 39% and Irish (Type 1) by up to 15%, compared to 11% for Singlish and 5% for Vietnamese (both Type 0). While this trend is by no means comprehensive, it suggests that effective use of MBERT requires considering the target language variety.

Our results thus support our hypotheses and give insight to the limitations of MBERT. Wordpieces appear in different contexts in different languages, and MBERT initially lacks enough exposure to wordpiece usage in Type 1/2 target languages to

[7] Our partition of the data is available at https://github.com/ethch18/parsing-mbert.

outperform baselines. However, increased exposure through additional language-specific pretraining can ameliorate this issue. Likewise, despite mBERT's attempt to balance its pretraining data, the existing vocabulary still favors languages that have been seen more. Augmenting the vocabulary can produce additional improvement for languages with greater proportions of unseen wordpieces. Overall, our findings are promising for low-resource language varieties, demonstrating that large improvements in performance are possible with the help of a little unlabeled data, and that the performance discrepancy of multilingual models for low-resource languages (Wu and Dredze, 2020) can be overcome.

5 Further Related Work

Our work builds on prior empirical studies on multilingual models, which probe the behavior and components of existing models to explain *why* they are effective. Cao et al. (2020), Pires et al. (2019), and Wu and Dredze (2019) note the importance of both vocabulary overlap and the relationship between languages in determining the effectiveness of multilingual models, but they primarily consider high-resource languages. On the other hand, Conneau et al. (2020b) and K et al. (2020) find vocabulary overlap to be less significant of a factor, instead attributing such models' successes to typological similarity and parameter sharing. Artetxe et al. (2020) emphasize the importance of sufficiently representing the target language in the vocabulary. Unlike these studies, we primarily consider *how* to improve the performance of multilingual models for a given target language variety. Though our experiments do not directly probe the impact of vocabulary overlap, we contribute further evaluation of the importance of improved modeling of the target variety.

Recent work has also proposed additional pretraining for general-purpose language models, especially with respect to domain (Alsentzer et al., 2019; Chakrabarty et al., 2019; Gururangan et al., 2020; Han and Eisenstein, 2019; Howard and Ruder, 2018; Logeswaran et al., 2019; Sun et al., 2019). Lakew et al. (2018) and Zoph et al. (2016) perform additional training on parallel data to adapt bilingual translation models to unseen target languages, while Mueller et al. (2020) improve a polyglot task-specific model by finetuning on labeled monolingual data in the target variety. To the best of our knowledge, our work is the first to demonstrate the effectiveness of additional pretraining for *massively multilingual* language models toward a target low-resource language variety, using only unlabeled data in the target variety.

6 Conclusion

We explore additional language-specific pretraining and vocabulary augmentation for multilingual contextual word representations in low-resource settings and find them to be effective for dependency parsing, especially in the lowest-resource cases. Our results demonstrate the significance of the relationship between a multilingual model's pretraining data and a target language. We expect that our findings can benefit practitioners in low-resource settings, and our data, code, and models are publicly available to accelerate further study.

Acknowledgments

We thank Jungo Kasai, Phoebe Mulcaire, members of UW NLP, and the anonymous reviewers for their helpful comments on preliminary versions of this paper. We also thank Hongmin Wang for providing the unlabeled Singlish data. This work was supported by a NSF Graduate Research Fellowship to LHL and by NSF grant 1813153.

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proc. of LREC*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proc. of Clinical Natural Language Processing Workshop*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proc. of ACL*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained language model for scientific text. In *Proc. of EMNLP*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proc. of ICLR*.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. In *Proc. of NAACL-HLT*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proc. of ACL*.

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*.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proc. of EMNLP-IJCNLP*.

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

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 Workshop for NLP Open Source Software (NLP-OSS)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proc. of EMNLP*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proc. of ACL*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *Proc. of ICLR*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proc. of EMNLP-IJCNLP*.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proc. of IWSLT*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs.CL].

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proc. of ACL*.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proc. of ACL*.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019a. Low-resource parsing with crosslingual contextualized representations. In *Proc. of CoNLL*.

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

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*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proc. of the First NLPL Workshop on Deep Learning for Natural Language Processing*.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? arXiv:1905.05583 [cs.CL].

Peter Trudgill. 2003. *A Glossary of Sociolinguistics*. Edinburgh University Press.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. In *Proc. of EMNLP-IJCNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Hongmin Wang, Jie Yang, and Yue Zhang. 2019. From genesis to creole language: Transfer learning for singlish universal dependencies parsing and pos tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(1).

Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial Singaporean English. In *Proc. of ACL*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proc. of EMNLP-IJCNLP*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proc. of the 5th Workshop on Representation Learning for NLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144 [cs.CL].

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, 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, Victoria Basmov, Colin Batchelor, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, 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, Maria Liovina, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Mackentanz, 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, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, 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, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proc. of EMNLP*.

A Supplementary Material to Accompany *Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank*

This supplement contains further details about the experiments presented in the main paper.

A.1 Vocabulary Augmentation and Statistics

Language	Original	Augmented
GA	12807	228
MT	49791	1124
SING	3	1
VI	6955	421

Table 3: Number of tokens with unknown wordpieces in the unlabeled dataset under original and augmented vocabularies.

We choose the vocabulary size to minimize the number of unknown wordpieces while maintaining a similar wordpiece-per-token ratio as the original MBERT vocabulary. Empirically, we find a vocabulary size of 5000 to best meet these criteria. Then, we tokenize the unlabeled data using both the new and original vocabularies. We compare the tokenizations of each word and note cases where the new vocabulary yields a tokenization with fewer unknown wordpieces than the original one. We select the 99 most common wordpieces that occur in these cases and use them to fill the 99 unused slots in MBERT's vocabulary. For Singlish, 99 such wordpieces are not available; we fill the remaining slots with the most common wordpieces in the new vocabulary.

Tab. 3 gives a comparison of the number of tokens with unknown wordpieces under the original and augmented MBERT vocabularies. The augmented vocabulary significantly decreases the number of unknowns, resulting in a specific embedding for most of the wordpieces.

A.2 Data Extraction and Preprocessing

In this section, we detail the steps used to obtain the pretraining data. After dataset-specific preprocessing, all datasets are tokenized with the multilingual spaCy tokenizer.[8] We then generate pretraining shards in a format acceptable by MBERT using scripts provided by Devlin et al. (2019) and the parameters listed in Tab. 7, which includes artificially

[8] https://spacy.io/models/xx

Language	Partition	# Sentences	# Tokens
GA	Train	858	20k
	Valid.	451	9.8k
	Test	454	10k
MT	Train	1123	23k
	Valid.	433	11k
	Test	518	10k
SING	Train	2465	22k
	Valid.	286	2.5k
	Test	299	2.7k
VI	Train	1400	24k
	Valid.	800	13k
	Test	800	14k

Table 4: Statistics for labeled Universal Dependencies datasets.

augmenting each dataset five times by masking different words with a probability of 0.15. Statistics for labeled datasets, which we use without modification, are provided in Tab. 4.

Wikipedia Data We draw data from the newest available Wikipedia dump[9] for the language at the time it was obtained: October 20, 2019 (Irish) and January 1, 2020 (Maltese, Vietnamese). We use WikiExtractor[10] to extract the article text, split sentences at periods, and remove the following items:

- Document start and end line
- Article titles and section headers
- Categories
- HTML content (e.g., `<br>`)

Articles are kept contiguous. The full Vietnamese Wikipedia consists of nearly 6.5 million sentences (141 million tokens); to simulate a truly low-resource setting, we randomly select 5% of the articles without replacement to use in our pretraining.

Singlish Data Beginning with the raw crawled sentences from Wang et al. (2017), we remove any sentences that appear verbatim in the validation or test sets of either their original treebank or our partition. Furthermore, we remove any sentences with fewer than five tokens or more than 50 tokens, as we observe that a large proportion of these sentences are either nonsensical or extended quotes from Standard English. We note that this dataset is non-contiguous: most sentences do not appear in a larger context.

A.3 Training Procedure

During pretraining, we use the original implementation of Devlin et al. (2019) but modify it to optimize based only on the masked language modeling (MLM) loss. Although Devlin et al. (2019) also trained on a next sentence prediction (NSP) loss, subsequent work has found joint optimization of NSP and MLM to be less effective than MLM alone (K et al., 2020; Lample and Conneau, 2019; Liu et al., 2019). Furthermore, in certain low-resource language varieties, fully contiguous data may not be available, rendering the NSP task ill-posed. We perform additional pretraining for up to 20 epochs, selecting our final model based on average validation LAS downstream.

Following prior work on parsing with MBERT (Kondratyuk and Straka, 2019), parsers are trained with a inverse square root learning rate decay and linear warmup, and gradual unfreezing and discriminative finetuning of the layers. These models are trained for up to 200 epochs with early stopping based on the validation performance. All parsers are implemented in AllenNLP, version 0.9.0 (Gardner et al., 2018).

Tab. 7 gives all hyperparameters kept constant during MBERT pretraining and parser training. The values for these hyperparameters largely reflect the defaults or recommendations specified in the implementations we used. For instance, the base learning rate for LAPT, VA, and TVA reflect recommendations in the code of Devlin et al. (2019), and the TVA embedding learning rate is equal to the learning rate used in the original pretraining of MBERT.

Due to the large number of parameters in MBERT, large batch sizes are sometimes infeasible. We reduce the batch size until training is able to complete succesfully on our GPU.

ELMO models are trained with the original implementation and default hyperparameter settings of Peters et al. (2018). However, following the implementation of Mulcaire et al. (2019b), we use a variable-length character vocabulary instead of a fixed-sized one to fully model the distribution in each language. FASTT is trained using the skipgram model for five epochs, with the default hyperparameters of Bojanowski et al. (2017). All experiments are variously conducted on a single

[9] `https://dumps.wikimedia.org/`

[10] `https://github.com/attardi/wikiextractor`

Representations	GA	MT	SING	VI
ELMO	10	10	5	10
LAPT	5	20	5	5
VA	10	15	1	5
TVA	15	20	20	5
LAPT + FT	20	10	1	5
VA + FT	10	10	1	5
TVA + FT	15	15	5	5

Table 5: Number of pretraining epochs used in final models, selected based on validation LAS scores.

Hyperparameter	Minimum	Maximum
Adam, Beta 1	0.9	0.9999
Adam, Beta 2	0.9	0.9999
Gradient Norm	1.0	10.0
Random Seed, Python	0	100000
Random Seed, Numpy	0	100000
Random Seed, PyTorch	0	100000

Table 6: Hyperparameter bounds for measuring variation.

NVIDIA Titan X or Titan XP GPU.

A.4 Hyperparameter Optimization

For our experiments, we fix both the pretraining and downstream architectures and tune only the number of pretraining epochs. For LAPT, VA, and TVA, we pretrain for an additional {1, 5, 10, 15, 20} epochs. For ELMO, we pretrain for {1, 3, 5, 10} epochs. Final selections are given in Tab. 5.

Measuring Variation We use Allentune (Dodge et al., 2019) to compute standard deviations for our experiments. For a given representation source, we randomly select five assignments of the following training hyperparameters via uniform sampling from the ranges specified in Tab. 6. To choose the best epoch for each method, we compute the average validation LAS for these five assignments to choose our final model. Then, we compute the average and standard deviation of the test LAS from each of these assignments.

In cases where a hyperparameter assignment yields exploding gradients and/or trends toward an infinite loss, we rerun the experiment to yield a feasible initialization.

Stage	**Hyperparameter**	**Value**
Data Creation	Max Sequence Length	128
	Max Predictions per Sequence	20
	Masked LM Probability	0.15
	Duplication Factor	5
Pretraining	Max Sequence Length	128
	Warmup Steps	1000
	Batch Size	{12, 16}
	Max Predictions per Sequence	20
	Masked LM Probability	0.15
	Learning Rate	0.00002
	TVA Embedding Learning Rate	0.0001
Parser	Dependency Arc Dimension	100
	Dependency Tag Dimension	100
	MBERT Layer Dropout	0.1
	ELMO Dropout	0.5
	Input Dropout	0.3
	Parser Dropout	0.3
	Optimizer	Adam
	Parser Learning Rate	0.001
	MBERT Learning Rate	0.00005
	Learning Rate Warmup Epochs	1
	Epochs	200
	Early Stopping (Patience)	20
	Batch Size	{8, 24, 64}
	BiLSTM Layers	3
	BiLSTM Hidden Size	400

Table 7: Hyperparameters for data creation, pretraining, and parser.