Extending Multilingual BERT to Low-Resource Languages

Zihan Wang[*]
University of Illinois Urbana-Champaign
Urbana, IL 61801, USA
zihanw2@illinois.edu

Karthikeyan K[*]
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh 208016, India
kkarthi@cse.iitk.ac.in

Stephen Mayhew[†]
Duolingo
Pittsburgh, PA, 15206, USA
stephen@duolingo.com

Dan Roth
University of Pennsylvania
Philadelphia, PA 19104, USA
danroth@seas.upenn.edu

Abstract

Multilingual BERT (M-BERT) has been a huge success in both supervised and zero-shot cross-lingual transfer learning. However, this success is focused only on the top 104 languages in Wikipedia it was trained on. In this paper, we propose a simple but effective approach to *extend* M-BERT (E-MBERT) so it can benefit any new language, and show that our approach aids languages that are already in M-BERT as well. We perform an extensive set of experiments with Named Entity Recognition (NER) on 27 languages, only 16 of which are in M-BERT, and show an average increase of about 6% F_1 on M-BERT languages and 23% F_1 increase on new languages. We release models and code at [1].

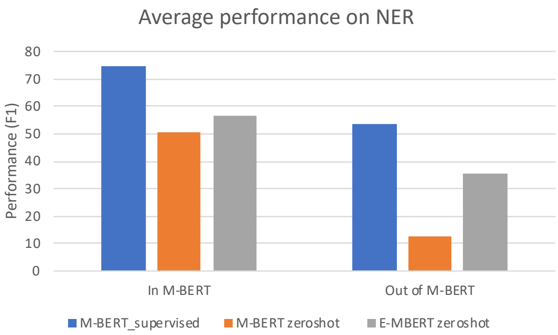

Figure 1: **Comparison between M-BERT and our proposed approach E-MBERT:** We report average zero-shot NER performance on 16 languages that are already *in* M-BERT and 11 new language that are *out* of M-BERT; M-BERT performance with supervised NER data is also reported as an upper-bound. In both languages in M-BERT and out of M-BERT, our method E-MBERT performs better than M-BERT.

1 Introduction

Recent works (Wu and Dredze, 2019; K et al., 2020) have shown the zero-shot cross-lingual ability of M-BERT (Devlin et al., 2018) on various semantic and syntactic tasks – just fine-tuning on English data allows the model to perform well on other languages. Cross-lingual learning is imperative for low-resource languages such as Somali and Uyghur, as obtaining supervised training data in these languages is particularly hard. However, M-BERT is not pre-trained with these languages, thus limiting its performance on them. Languages like Oromo, Hausa, Amharic and Akan are spoken by more than 20 million people, yet M-BERT does not cover them. Indeed, there are about 4000[2] written languages, of which M-BERT covers only the top 104 languages (less than 3%).

One straightforward way to extend the notion of M-BERT to languages not covered by it is to train a new M-BERT from scratch to include the new language. However, this is extremely time-consuming and expensive: training BERT-base takes about four days with four cloud TPUs (Devlin et al., 2019). Alternatively, one can train a BERT with two languages, a high resource one (typically English) and the target, low resource language. This is also known as Bilingual BERT (B-BERT) (K et al., 2020), which is more efficient than M-BERT. However, one major disadvantage of B-BERT is that we can not make use of data from related languages.

To accommodate a language not in M-BERT, we propose an efficient approach, EXTEND. EXTEND works by first enlarging the vocabulary of M-BERT to accommodate the new language and then continuing pre-training on this language. Our approach trains for less than 7 hours on a single cloud TPU.

We perform comprehensive experiments

[*] Equal Contribution; most of this work was done while the authors interned at the University of Pennsylvania.

[†] This work was done while the author was a student at the University of Pennsylvania.

[1] http://cogcomp.org/page/publication_view/912

[2] https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2649–2656
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

on cross-lingual NER on the LORELEI dataset (Strassel and Tracey, 2016) with 27 languages of which 11 languages are not present in M-BERT. As shown in Figure 1, our approach significantly outperforms M-BERT when the target language is not in the 104 languages in M-BERT and it is superior to M-BERT even for the high-resource languages that are already in it.

The key contributions of our work are (i) EXTEND, a simple yet novel approach to add a new language to M-BERT, (ii) experiments that show EXTEND improves M-BERT for languages that are in M-BERT as well as those that are not, (iii) results showing that EXTEND provides performance and efficiency improvements, in most cases, over B-BERT.

2 Related works

Cross-lingual learning has seen increased interest in NLP, with such works as BiCCA (Faruqui and Dyer, 2014), LASER (Artetxe and Schwenk, 2019) and XLM (Conneau and Lample, 2019). Although these models have been successful, they need cross-lingual supervision such as bilingual dictionaries or parallel corpora (Upadhyay et al., 2016), which are particularly challenging to obtain for low-resource languages. Our work differs in that we do not require such supervision. While other approaches like MUSE (Lample et al., 2018) and VecMap (Artetxe et al., 2018) can work without any cross-lingual supervision, M-BERT alone often outperforms these approaches (K et al., 2020).

Schuster et al. (2019) has a continuing training setting that is similar to ours. However, their approach focuses on comparing between whether B-BERT (JointPair) learns cross-lingual features from overlapping word-pieces, while ours aims at improving M-BERT on target languages, and addresses the problem of missing word-pieces. We show that our EXTEND method works well on M-BERT, and is better than B-BERT in several languages, whereas their method (MonoTrans) has a similar performance as B-BERT. This implies that our EXTEND method benefits from the multilinguality of the base model (M-BERT vs BERT).

A recent work on multilingual BERT (Wu and Dredze, 2020) reveals that a monolingual BERT underperforms multilingual BERT on low-resource cases. Our work also identifies this phenomenon in some languages (see Appendix), and we then present an effective way of extending M-BERT to work even better than multilingual BERT on these low-resource languages.

3 Background

3.1 Multilingual BERT (M-BERT)

M-BERT is a transformer language model pretrained with Wikipedia text of the top 104 languages in Wikipedia. M-BERT uses the same pretraining objectives as BERT – masked language model and next sentence prediction (Devlin et al., 2019) – and is surprisingly cross-lingual despite not being trained with any cross-lingual objective or aligned data. For cross-lingual transfer, M-BERT is fine-tuned on supervised data in high-resource languages and tested on the target language.

3.2 Bilingual BERT (B-BERT)

B-BERT is trained in the same way as M-BERT except that it contains only two languages – English and the target language. Recent works have shown the effectiveness of M-BERT (Pires et al., 2019; Wu and Dredze, 2019), and B-BERT (K et al., 2020) on NER and other tasks.

4 Our Method: Extend

In this section, we discuss our training protocol EXTEND which incorporates the target language by extending the vocabulary, encoders and decoders, and then continues pre-training.

Let M-BERT's vocabulary be V_{mbert} and let the extended new vocabulary be V_{new}. Throughout the paper, we fix the size of $|V_{new}| = 30,000$. The training goes as following:

1. Extend the vocabulary, encoder, and decoder to accommodate V_{new}. That is, let $|V_{extra}| = |V_{new} - V_{mbert}|$, and increase the dimension of size $|V_{mbert}|$ to $|V_{mbert}| + |V_{extra}|$.
2. Initialize all the new weights with M-BERT's default weight initialization.
3. Continue pre-training with monolingual data of the target language. We call the trained model E-MBERT.

5 Experiments

The goal of our experiments is to establish the extent to which our method EXTEND and the resulting model E-MBERT (i) improves over M-BERT, (ii) does not necessary require additional monolingual data to continue training on, and (iii) is both effective and efficient compared to B-BERT.

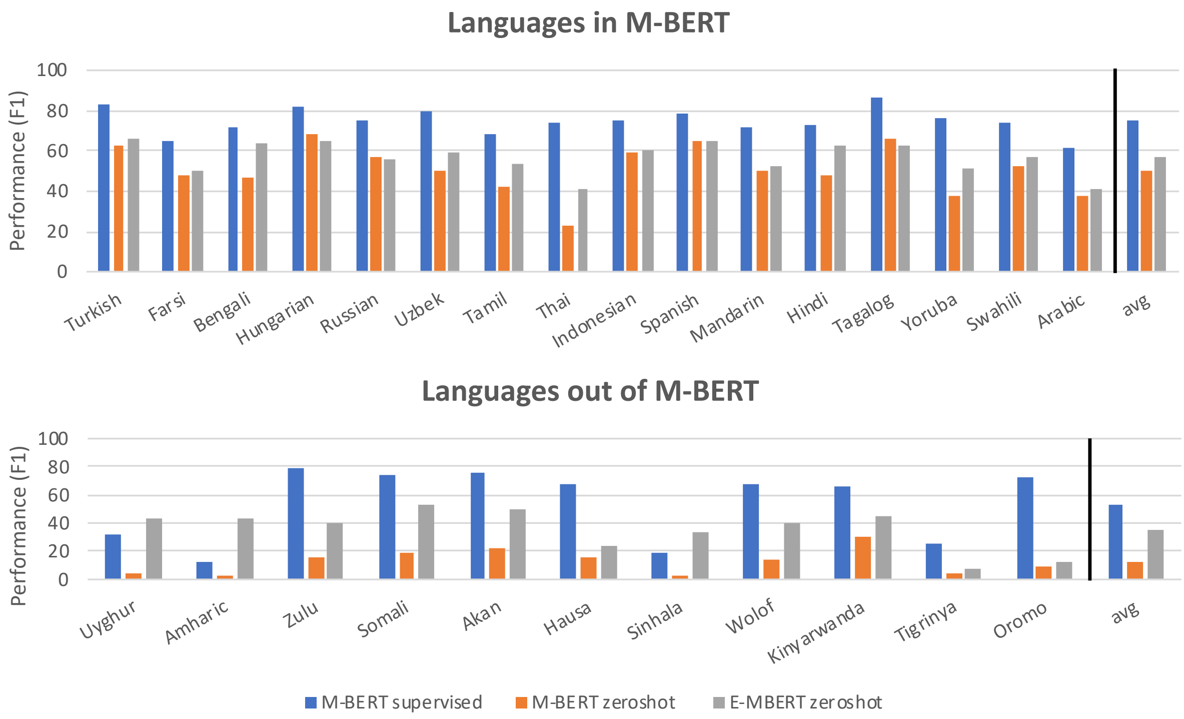

Figure 2: **Comparison between M-BERT and E-MBERT:** We compare zero-shot cross-lingual NER performance on M-BERT and E-MBERT using 27 languages. The languages are ordered left to right by amount of monolingual text data in LORELEI. Whether the languages are in or out of M-BERT, E-MBERT performs better.

5.1 Experimental Settings

Dataset. Our text corpus and NER dataset are from LORELEI, preprocessed using the tokenization method from BERT. For zero-shot cross-lingual NER, we evaluate the performance on the whole annotated set; for supervised learning, since we just want an understanding of an upper bound, we apply cross validation to estimate the performance: each fold is evaluated by a model trained on the other folds, and the average F_1 is reported.

NER Model. We use AllenNLP (Gardner et al., 2018) with a standard Bi-LSTM-CRF (Ma and Hovy, 2016; Lample et al., 2016) framework. The score reported in NER is the F_1 score averaged across five runs with different random seeds.

BERT training. While extending, we use a batch size of 32 and a learning rate of 2e-5, and train for 500K iterations. Whereas for B-BERT we use a batch size of 32 and learning rate of 1e-4 and train for 2M iterations. We follow BERT's setting for all other hyperparameters.

5.2 Comparing E-MBERT and M-BERT

We compare the cross-lingual zero-shot NER performance of M-BERT and E-MBERT. We train with supervised English NER data and report the performance on the target language. We also report the performance when there is supervision on the target language as a reasonable "upper-bound" on the dataset. From Figure 2, we can see that in almost all languages, EXTEND brings a performance improvement irrespective of whether or not the language exists in M-BERT.

It is clear that using EXTEND, the model performs better when the language is not already present; however, it is intriguing that E-MBERT improves when the language is already present. We attribute this to three reasons:

- Increased vocabulary size of target language. Since most languages have a significantly smaller dataset than English, they have a smaller vocabulary in M-BERT; our approach eliminates this issue. Note that it is infeasible to train single M-BERT with larger vocabulary sizes for every language, as this will create a vast vocabulary.
- Extra monolingual data – more monolingual data in the target language can be beneficial.
- E-MBERT is more focused on the target language, as during the last 500K steps, it is optimized to perform well on it.

5.3 Extra vocabulary

To address the possibility of out-of-vocabulary word-pieces (e.g. a new script), we enlarged the vo-

Lang	M-BERT	E w/ LRL	E w/ Wiki
Russian	56.56	55.70	56.64
Thai	22.46	40.99	38.35
Hindi	48.31	62.72	62.77

Table 1: **Performance of EXTEND with different number of new vocabulary introduced:** a larger vocabulary in general performs better than using the M-BERT version. However, even without adding new vocabulary, EXTEND still improves the performance of the model.

cabulary of M-BERT by the vocabulary estimated from the new corpus. From Table 1, it is clear that a larger vocabulary helps the models a lot. It is also noteworthy to point out that even without introducing this new vocabulary, the continue training framework can still familiarize the model with the new data, and thus bringing up the performance.

Lang	M-BERT	E w/ LRL	E w/ Wiki
Russian	56.56	55.70	56.64
Thai	22.46	40.99	38.35
Hindi	48.31	62.72	62.77

Table 2: **Performance of M-BERT, EXTEND with LORELEI data and EXTEND with Wikipedia data:** Even without the additional data from LORELEI (LRL), our EXTEND method works comparably well.

5.4 Extra data

The effectiveness of E-MBERT may be attributed to the extra monolingual data introduced. To explore the performance of E-MBERT without this extra training data, we EXTEND with Wikipedia data, which is already used in M-BERT, while controling all other settings to be the same. From Table 2, we can see that even without additional data, E-MBERT's performance does not degrade.

5.5 Comparing E-MBERT and B-BERT

Another way of addressing M-BERT on unseen languages is to train B-BERT on source and target. Both E-MBERT and B-BERT use the same text corpus in the target language; for the source, we use subsampled English Wikipedia data. We focus only on languages that are not in M-BERT so that E-MBERT will not have an advantage on the target language because of Wikipedia data. Although the English corpus of the two models are different, the difference is marginal considering its size. Indeed we show that B-BERT and E-MBERT have similar performance on English NER (see Appendix).

Lang	B-BERT	EXTEND
Somali	51.18	**53.63**
Amharic	38.66	**43.70**
Uyghur	21.94	**42.98**
Akan	48.00	**49.02**
Hausa	**26.45**	24.37
Wolof	**39.92**	39.70
Zulu	**44.08**	39.65
Tigrinya	6.34	**7.61**
Oromo	8.45	**12.28**
Kinyarwanda	**46.72**	44.40
Sinhala	16.93	**33.97**
Average	31.70	**35.57**

Table 3: **Comparison between B-BERT and E-MBERT:** We compare B-BERT vs E-MBERT training protocols. Both models use same target language monolingual data. E-MBERT performs better than B-BERT in more languages and in average.

From Table 3, we can see that E-MBERT often outperforms B-BERT. Moreover, B-BERT is trained for 2M steps for convergence, while E-MBERT requires only 500k steps. We believe that this advantage comes for the following reason: E-MBERT makes use of a multilingual model, which potentially contains similar languages that help transfer knowledge from English to target, while B-BERT can only leverage English data. For example, in the case of Sinhala and Uyghur, a comparatively high-resource related language like Tamil and Turkish in M-BERT can help E-MBERT learn the target language better.

5.6 Rate of Convergence

In this subsection, we study the convergence rate of E-MBERT and B-BERT. We evaluate these two models on two languages, Hindi (in M-BERT) and Sinhala (not in M-BERT), and report the results in Figure 3. We can see that E-MBERT is able to converge within just 100K steps, while B-BERT takes more than 1M steps to converge. This shows that E-MBERT is much more efficient than B-BERT.

5.7 Performance on non-target languages

The EXTEND methods results in focusing the base model on the target language, and this degrades performance on the other languages that are not the target language. We report the performance of the Hindi and Sinhala E-MBERT models evaluated on the other languages in the Appendix.

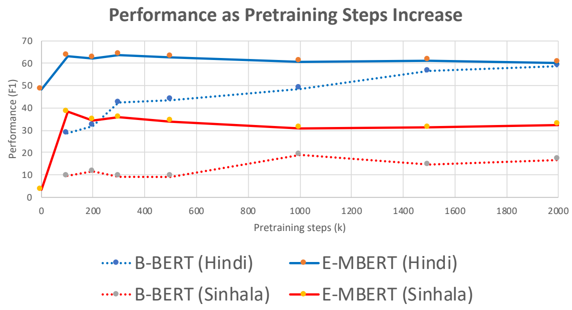

Figure 3: **Performance of B-BERT and E-MBERT as number of pre-training steps increases:** E-MBERT converges in 100K steps, which is 1/10 of B-BERT.

6 Conclusions and Future work

We proposed EXTEND, an efficient method that extends M-BERT to deal with languages that were originally outside it. Our method has shown greatly improved performance across several languages comparing to M-BERT and B-BERT.

While EXTEND deals with one language each time, it would be an interesting future work to extend on multiple languages at the same time. Furthermore, instead of randomly initializing the embeddings of a new vocabulary, we could possibly use alignment models like MUSE or VecMap with bilingual dictionaries to initialize. We could also try to apply our approach to better models like RoBERTa (Liu et al., 2019).

Acknowledgments

This work was supported by contracts FA8750-19-2-1004, HR0011-15-C-0113, and HR0011-18-2-0052 from the US Defense Advanced Research Projects Agency (DARPA), by Google Cloud, and by Cloud TPUs from Google's TensorFlow Research Cloud (TFRC). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government or Google.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Multilingual bert - r.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

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 *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

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 preprint arXiv:1907.11692*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3273–3280.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. In *ACL*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

A Appendices

A.1 Performance of E-MBERT on English:

The knowledge of E-MBERT on English (source language) is not affected. From Table 4, we can see that, except for few languages, the English performance of E-MBERT is almost as good as M-BERT's.

A.2 Detailed data on all languages

In Table 5, we report the full result on comparing M-BERT and E-MBERT.

We can also see that EXTEND is not only useful for cross-lingual performance but also for useful for supervised performance (in almost all cases).

We also notice that extending on one language hurts the transferability to other languages.

A.3 Comparison between B-BERT and E-MBERT:

In Table 6 we reported the performance of EXTEND and B-BERT on both English as well as target. We can see that English performance of B-BERT is mostly better than EXTEND. However, in most cases EXTEND performs better on target language. This indicates that E-MBERT does not have an unfair advantage on English.

EXTEND Language	**E M-BERT**
OUT OF BERT	
Akan	79.19
Amharic	78.36
Hausa	74.24
Somali	78.6
Wolof	78.11
Zulu	79.32
Uyghur	77.76
Tigrinya	76.21
Oromo	76.06
Kinyarwanda	73.05
Sinhala	73.7
IN BERT	
Arabic	77.67
Bengali	76.2
Mandarin	78.58
Farsi	77.57
Hindi	78.86
Hungarian	78.92
Indonesian	80.93
Russian	80.87
Spanish	81.15
Swahili	77.72
Tamil	77.6
Tagalog	79.56
Thai	78.21
Turkish	79.49
Uzbek	77.19
Yoruba	77.55
M-BERT	79.37

Table 4: **Performance on English:** We report the English NER performance of M-BERT as well as performance E-MBERT.

In BERT								
Model	**M-sup**	**M-zero**	**E-sup**	**E-zero**	Hindi	Sinhala	Corpus (M)	NER (k)
Arabic	61.14	37.56	61.97	40.83	19.2	16.72	0.19	5.50
Bengali	71.29	46.18	84.44	63.49	17.94	14.01	10.19	11.65
Mandarin	71.76	50.0	73.86	52.30	8.88	24.64	1.66	8.05
Farsi	65.09	47.71	68.27	50.26	22.38	20.44	10.32	4.38
Hindi	72.88	48.31	81.15	62.72	62.72	18.0	1.66	6.22
Hungarian	81.98	68.26	82.08	64.36	24.38	35.74	10.09	5.81
Indonesian	75.67	58.91	80.09	60.73	29.5	37.89	1.75	6.96
Russian	75.60	56.56	76.51	55.70	26.08	36.15	10.07	7.26
Spanish	78.12	64.53	78.14	64.75	37.06	47.32	1.68	3.48
Swahili	74.26	52.39	81.9	57.21	25.46	31.91	0.29	5.61
Tamil	68.55	41.68	77.91	53.42	14.75	12.96	4.47	15.51
Tagalog	85.98	66.50	88.63	62.61	34.73	42.16	0.33	6.98
Thai	73.58	22.46	86.40	40.99	4.03	3.78	4.47	15.51
Turkish	82.55	62.80	87.02	66.19	34.34	39.23	10.39	7.09
Uzbek	79.36	49.56	84.79	59.68	21.84	28.83	4.91	11.82
Yoruba	75.75	37.13	81.34	50.72	19.14	25.04	0.30	3.21
Out of BERT								
Akan	75.87	21.96	79.33	49.02	12.82	35.2	0.52	8.42
Amharic	11.79	3.27	79.09	43.70	3.95	3.9	1.70	5.48
Hausa	67.67	15.36	75.73	24.37	12.58	14.77	0.19	5.64
Somali	74.29	18.35	84.56	53.63	15.84	21.64	0.60	4.16
Wolof	67.10	13.63	70.27	39.70	9.83	26.45	0.09	10.63
Zulu	78.89	15.82	84.50	39.65	12.3	13.72	0.92	11.58
Uyghur	32.64	3.59	79.94	42.98	1.45	1.52	1.97	2.45
Tigrinya	24.75	4.74	79.42	7.61	7.91	5.71	0.01	2.20
Oromo	72.00	9.34	72.78	12.28	6.84	10.11	0.01	2.96
Kinyarwanda	65.85	30.18	74.46	44.40	26.55	32.3	0.06	0.95
Sinhala	18.12	3.43	71.63	33.97	3.39	33.97	0.10	1.02

Table 5: In the order from left to right, column means: M-BERT with supervision, M-BERT zero-shot cross-lingual, E-MBERT with supervision, E-MBERT zero-shot cross-lingual. Then we give performance of Hindi and Sinhala E-MBERT models when evaluated on all the languages. The last two columns are dataset statistics, with number of million lines in the LORELEI corpus and number of thousand lines in LORELEI NER dataset.

	English		Target	
Language	**E-MBERT**	**B-BERT**	**E-MBERT**	**B-BERT**
Akan	79.19	77.49	49.02	48.00
Amharic	78.36	78.44	43.70	38.66
Hausa	74.24	80.13	24.37	26.45
Somali	78.60	79.17	53.63	51.18
Wolof	78.11	81.01	39.70	39.92
Zulu	79.32	81.82	39.65	44.08
Uyghur	77.76	79.65	42.98	21.94
Tigrinya	76.21	80.35	7.61	6.34
Oromo	76.06	78.13	12.28	8.45
Kinyarwanda	73.05	79.37	44.4	46.72
Sinhal	73.70	80.04	33.97	16.93

Table 6: **Comparison Between B-BERT vs E-MBERT:** We compare the performance of E-MBERT with B-BERT on both English and target language. As a reference, performance of M-BERT is 79.37 on English. This shows that neither B-BERT nor E-MBERT gets unfair advantage from the English part of the model.