How multilingual is Multilingual BERT?

Telmo Pires* **Eva Schlinger** **Dan Garrette**
Google Research
{telmop,eschling,dhgarrette}@google.com

Abstract

In this paper, we show that Multilingual BERT (M-BERT), released by Devlin et al. (2019) as a single language model pre-trained from monolingual corpora in 104 languages, is surprisingly good at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. To understand why, we present a large number of probing experiments, showing that transfer is possible even to languages in different scripts, that transfer works best between typologically similar languages, that monolingual corpora can train models for code-switching, and that the model can find translation pairs. From these results, we can conclude that M-BERT does create multilingual representations, but that these representations exhibit systematic deficiencies affecting certain language pairs.

1 Introduction

Deep, contextualized language models provide powerful, general-purpose linguistic representations that have enabled significant advances among a wide range of natural language processing tasks (Peters et al., 2018b; Devlin et al., 2019). These models can be pre-trained on large corpora of readily available unannotated text, and then fine-tuned for specific tasks on smaller amounts of supervised data, relying on the induced language model structure to facilitate generalization beyond the annotations. Previous work on model probing has shown that these representations are able to encode, among other things, syntactic and named entity information, but they have heretofore focused on what models trained on English capture about English (Peters et al., 2018a; Tenney et al., 2019b,a).

In this paper, we empirically investigate the degree to which these representations generalize *across* languages. We explore this question using Multilingual BERT (henceforth, M-BERT), released by Devlin et al. (2019) as a single language model pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages.[1] M-BERT is particularly well suited to this probing study because it enables a very straightforward approach to zero-shot cross-lingual model transfer: we fine-tune the model using task-specific supervised training data from one language, and evaluate that task in a different language, thus allowing us to observe the ways in which the model generalizes information across languages.

Our results show that M-BERT is able to perform cross-lingual generalization surprisingly well. More importantly, we present the results of a number of probing experiments designed to test various hypotheses about how the model is able to perform this transfer. Our experiments show that while high lexical overlap between languages improves transfer, M-BERT is also able to transfer between languages written in different scripts—thus having *zero* lexical overlap—indicating that it captures multilingual representations. We further show that transfer works best for typologically similar languages, suggesting that while M-BERT's multilingual representation is able to map learned structures onto new vocabularies, it does not seem to learn systematic transformations of those structures to accommodate a target language with different word order.

2 Models and Data

Like the original English BERT model (henceforth, EN-BERT), M-BERT is a 12 layer transformer (Devlin et al., 2019), but instead of be-

*Google AI Resident.

[1] https://github.com/google-research/bert

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001
Florence, Italy, July 28 - August 2, 2019.

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Table 2: POS accuracy on a subset of UD languages.

ing trained only on monolingual English data with an English-derived vocabulary, it is trained on the Wikipedia pages of 104 languages with a shared word piece vocabulary. It does not use any marker denoting the input language, and does not have any explicit mechanism to encourage translation-equivalent pairs to have similar representations.

For NER and POS, we use the same sequence tagging architecture as Devlin et al. (2019). We tokenize the input sentence, feed it to BERT, get the last layer's activations, and pass them through a final layer to make the tag predictions. The whole model is then fine-tuned to minimize the cross entropy loss for the task. When tokenization splits words into multiple pieces, we take the prediction for the first piece as the prediction for the word.

2.1 Named entity recognition experiments

We perform NER experiments on two datasets: the publicly available CoNLL-2002 and -2003 sets, containing Dutch, Spanish, English, and German (Tjong Kim Sang, 2002; Sang and Meulder, 2003); and an in-house dataset with 16 languages,[2] using the same CoNLL categories. Table 1 shows M-BERT zero-shot performance on all language pairs in the CoNLL data.

2.2 Part of speech tagging experiments

We perform POS experiments using Universal Dependencies (UD) (Nivre et al., 2016) data for 41 languages.[3] We use the evaluation sets from Zeman et al. (2017). Table 2 shows M-BERT zero-shot results for four European languages. We see that M-BERT generalizes well across languages, achieving over 80% accuracy for all pairs.

[2]Arabic, Bengali, Czech, German, English, Spanish, French, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Turkish, and Chinese.

[3]Arabic, Bulgarian, Catalan, Czech, Danish, German, Greek, English, Spanish, Estonian, Basque, Persian, Finnish, French, Galician, Hebrew, Hindi, Croatian, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Marathi, Dutch, Norwegian (Bokmaal and Nynorsk), Polish, Portuguese (European and Brazilian), Romanian, Russian, Slovak, Slovenian, Swedish, Tamil, Telugu, Turkish, Urdu, and Chinese.

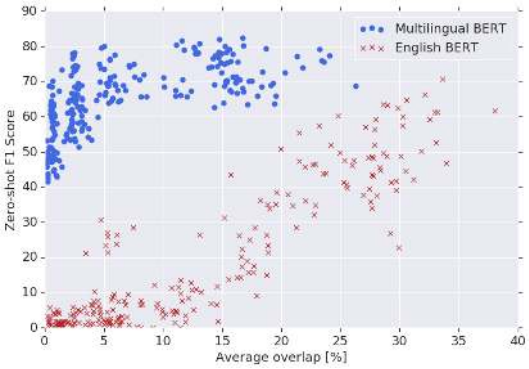

Figure 1: Zero-shot NER F1 score versus entity word piece overlap among 16 languages. While performance using EN-BERT depends directly on word piece overlap, M-BERT's performance is largely independent of overlap, indicating that it learns multilingual representations deeper than simple vocabulary memorization.

3 Vocabulary Memorization

Because M-BERT uses a single, multilingual vocabulary, one form of cross-lingual transfer occurs when word pieces present during fine-tuning also appear in the evaluation languages. In this section, we present experiments probing M-BERT's dependence on this superficial form of generalization: How much does transferability depend on lexical overlap? And is transfer possible to languages written in different scripts (*no* overlap)?

3.1 Effect of vocabulary overlap

If M-BERT's ability to generalize were mostly due to vocabulary memorization, we would expect zero-shot performance on NER to be highly dependent on word piece overlap, since entities are often similar across languages. To measure this effect, we compute E_{train} and E_{eval}, the sets of word pieces used in entities in the training and evaluation datasets, respectively, and define overlap as the fraction of common word pieces used in the entities: $overlap = |E_{train} \cap E_{eval}| \; / \; |E_{train} \cup E_{eval}|$.

Figure 1 plots NER F1 score versus entity overlap for zero-shot transfer between every language pair in an in-house dataset of 16 languages, for both M-BERT and EN-BERT.[4] We can see that

[4]Results on CoNLL data follow the same trends, but those trends are more apparent with 16 languages than with 4.

Model	EN	DE	NL	ES
Lample et al. (2016)	90.94	78.76	81.74	85.75
EN-BERT	91.07	73.32	84.23	81.84

Table 3: NER F1 results fine-tuning and evaluating on the *same* language (not zero-shot transfer).

	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

performance using EN-BERT depends directly on word piece overlap: the ability to transfer deteriorates as word piece overlap diminishes, and F1 scores are near zero for languages written in different scripts. M-BERT's performance, on the other hand, is flat for a wide range of overlaps, and even for language pairs with almost no lexical overlap, scores vary between 40% and 70%, showing that M-BERT's pretraining on multiple languages has enabled a representational capacity deeper than simple vocabulary memorization.[5]

To further verify that EN-BERT's inability to generalize is due to its lack of a multilingual representation and not an inability of its English-specific word piece vocabulary to represent data in other languages, we evaluate on *non*-cross-lingual NER and see that it performs comparably to a previous state of the art model (see Table 3).

3.2 Generalization across scripts

M-BERT's ability to transfer between languages that are written in different scripts, and thus have effectively *zero* lexical overlap, is surprising given that it was trained on separate monolingual corpora and not with a multilingual objective. To probe deeper into how the model is able to perform this generalization, Table 4 shows a sample of POS results for transfer across scripts.

Among the most surprising results, an M-BERT model that has been fine-tuned using only POS-labeled Urdu (written in Arabic script), achieves 91% accuracy on Hindi (written in Devanagari script), even though it has never seen a single POS-tagged Devanagari word. This provides clear evidence of M-BERT's multilingual representation ability, mapping structures onto new vocabularies based on a shared representation induced solely from monolingual language model training data.

However, cross-script transfer is less accurate for other pairs, such as English and Japanese, indicating that M-BERT's multilingual representation is not able to generalize equally well in all cases. A possible explanation for this, as we will see in section 4.2, is typological similarity. English and Japanese have a different order of subject, verb and object, while English and Bulgarian have the same, and M-BERT may be having trouble generalizing across different orderings.

[5]Individual language trends are similar to aggregate plots.

4 Encoding Linguistic Structure

In the previous section, we showed that M-BERT's ability to generalize cannot be attributed solely to vocabulary memorization, and that it must be learning a deeper multilingual representation. In this section, we present probing experiments that investigate the nature of that representation: How does typological similarity affect M-BERT's ability to generalize? Can M-BERT generalize from monolingual inputs to code-switching text? Can the model generalize to transliterated text without transliterated language model pretraining?

4.1 Effect of language similarity

Following Naseem et al. (2012), we compare languages on a subset of the WALS features (Dryer and Haspelmath, 2013) relevant to grammatical ordering.[6] Figure 2 plots POS zero-shot accuracy against the number of common WALS features. As expected, performance improves with similarity, showing that it is easier for M-BERT to map linguistic structures when they are more similar, although it still does a decent job for low similarity languages when compared to EN-BERT.

4.2 Generalizing across typological features

Table 5 shows macro-averaged POS accuracies for transfer between languages grouped according to two typological features: subject/object/verb order, and adjective/noun order[7] (Dryer and Haspelmath, 2013). The results reported include only zero-shot transfer, i.e. they do not include cases

[6]81A (Order of Subject, Object and Verb), 85A (Order of Adposition and Noun), 86A (Order of Genitive and Noun), 87A (Order of Adjective and Noun), 88A (Order of Demonstrative and Noun), and 89A (Order of Numeral and Noun).

[7]**SVO languages**: Bulgarian, Catalan, Czech, Danish, English, Spanish, Estonian, Finnish, French, Galician, Hebrew, Croatian, Indonesian, Italian, Latvian, Norwegian (Bokmaal and Nynorsk), Polish, Portuguese (European and Brazilian), Romanian, Russian, Slovak, Slovenian, Swedish, and Chinese. **SOV Languages**: Basque, Farsi, Hindi, Japanese, Korean, Marathi, Tamil, Telugu, Turkish, and Urdu.

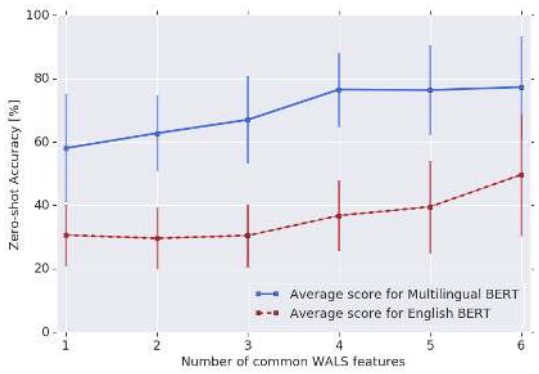

Figure 2: Zero-shot POS accuracy versus number of common WALS features. Due to their scarcity, we exclude pairs with no common features.

	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

Table 5: Macro-average POS accuracies when transferring between SVO/SOV languages or AN/NA languages. Row = fine-tuning, column = evaluation.

training and testing on the same language. We can see that performance is best when transferring between languages that share word order features, suggesting that while M-BERT's multilingual representation is able to map learned structures onto new vocabularies, it does not seem to learn systematic transformations of those structures to accommodate a target language with different word order.

4.3 Code switching and transliteration

Code-switching (CS)—the mixing of multiple languages within a single utterance—and transliteration—writing that is not in the language's standard script—present unique test cases for M-BERT, which is pre-trained on monolingual, standard-script corpora. Generalizing to code-switching is similar to other cross-lingual transfer scenarios, but would benefit to an even larger degree from a shared multilingual representation. Likewise, generalizing to transliterated text is similar to other cross-script transfer experiments, but has the additional caveat that M-BERT was not pre-trained on text that looks like the target.

We test M-BERT on the CS Hindi/English UD corpus from Bhat et al. (2018), which provides texts in two formats: *transliterated*, where Hindi words are written in Latin script, and *corrected*, where annotators have converted them back to Devanagari script. Table 6 shows the results for models fine-tuned using a combination of monolingual Hindi and English, and using the CS training set (both fine-tuning on the script-corrected version of the corpus as well as the transliterated version).

	Corrected	Transliterated
Train on monolingual HI+EN		
M-BERT	86.59	50.41
Ball and Garrette (2018)	—	77.40
Train on code-switched HI/EN		
M-BERT	90.56	85.64
Bhat et al. (2018)	—	90.53

Table 6: M-BERT's POS accuracy on the code-switched Hindi/English dataset from Bhat et al. (2018), on script-corrected and original (transliterated) tokens, and comparisons to existing work on code-switch POS.

For script-corrected inputs, i.e., when Hindi is written in Devanagari, M-BERT's performance when trained only on monolingual corpora is comparable to performance when training on code-switched data, and it is likely that some of the remaining difference is due to domain mismatch. This provides further evidence that M-BERT uses a representation that is able to incorporate information from multiple languages.

However, M-BERT is not able to effectively transfer to a transliterated target, suggesting that it is the language model pre-training on a particular language that allows transfer to that language. M-BERT is outperformed by previous work in both the monolingual-only and code-switched supervision scenarios. Neither Ball and Garrette (2018) nor Bhat et al. (2018) use contextualized word embeddings, but both incorporate explicit transliteration signals into their approaches.

5 Multilingual characterization of the feature space

In this section, we study the structure of M-BERT's feature space. If it is multilingual, then the transformation mapping between the same sentence in 2 languages should not depend on the sentence itself, just on the language pair.

5.1 Experimental Setup

We sample 5000 pairs of sentences from WMT16 (Bojar et al., 2016) and feed each sentence (separately) to M-BERT with no fine-tuning. We then extract the hidden feature activations at each layer for each of the sentences, and average the representations for the input tokens except [CLS] and [SEP], to get a vector for each sentence, at each layer l, $v^{(l)}_{\text{LANG}}$. For each pair of sentences,

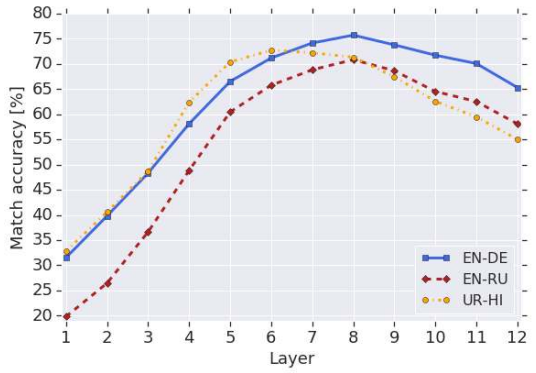

Figure 3: Accuracy of nearest neighbor translation for EN-DE, EN-RU, and HI-UR.

e.g. $(v^{(l)}_{\text{EN}_i}, v^{(l)}_{\text{DE}_i})$, we compute the vector pointing from one to the other and average it over all pairs: $\bar{v}^{(l)}_{\text{EN}\to\text{DE}} = \frac{1}{M}\sum_i \left(v^{(l)}_{\text{DE}_i} - v^{(l)}_{\text{EN}_i}\right)$, where M is the number of pairs. Finally, we translate each sentence, $v^{(l)}_{\text{EN}_i}$, by $\bar{v}^{(l)}_{\text{EN}\to\text{DE}}$, find the closest German sentence vector[8], and measure the fraction of times the nearest neighbour is the correct pair, which we call the "nearest neighbor accuracy".

5.2 Results

In Figure 3, we plot the nearest neighbor accuracy for EN-DE (solid line). It achieves over 50% accuracy for all but the bottom layers,[9] which seems to imply that the hidden representations, although separated in space, share a common subspace that represents useful linguistic information, in a language-agnostic way. Similar curves are obtained for EN-RU, and UR-HI (in-house dataset), showing this works for multiple languages.

As to the reason why the accuracy goes down in the last few layers, one possible explanation is that since the model was pre-trained for language modeling, it might need more language-specific information to correctly predict the missing word.

6 Conclusion

In this work, we showed that M-BERT's robust, often surprising, ability to generalize cross-lingually is underpinned by a multilingual representation, without being explicitly trained for it. The model handles transfer across scripts and to code-switching fairly well, but effective transfer to typologically divergent and transliterated targets will likely require the model to incorporate an explicit multilingual training objective, such as that used by Lample and Conneau (2019) or Artetxe and Schwenk (2018).

As to why M-BERT generalizes across languages, we hypothesize that having word pieces used in all languages (numbers, URLs, etc) which have to be mapped to a shared space forces the co-occurring pieces to also be mapped to a shared space, thus spreading the effect to other word pieces, until different languages are close to a shared space.

It is our hope that these kinds of probing experiments will help steer researchers toward the most promising lines of inquiry by encouraging them to focus on the places where current contextualized word representation approaches fall short.

[8] In terms of ℓ_2 distance.

[9] Our intuition is that the lower layers have more "token level" information, which is more language dependent, particularly for languages that share few word pieces.

7 Acknowledgements

We would like to thank Mark Omernick, Livio Baldini Soares, Emily Pitler, Jason Riesa, and Slav Petrov for the valuable discussions and feedback.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of EMNLP*.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal dependency parsing for Hindi-English code-switching. In *Proceedings of NAACL*.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*.

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

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://wals.info/.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016.

Neural architectures for named entity recognition. In *Proceedings of NAACL*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018a. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of EMNLP*.

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

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

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

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of ICLR*.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of CoNLL*.

A Model Parameters

All models were fine-tuned with a batch size of 32, and a maximum sequence length of 128 for 3 epochs. We used a learning rate of $3e{-}5$ with learning rate warmup during the first 10% of steps, and linear decay afterwards. We also applied 10% dropout on the last layer. No parameter tuning was performed. We used the `BERT-Base, Multilingual Cased` checkpoint from https://github.com/google-research/bert.

B CoNLL Results for EN-BERT

Fine-tuning \Eval	EN	DE	NL	ES
EN	**91.07**	24.38	40.62	49.99
DE	55.36	**73.32**	54.84	50.80
NL	59.36	27.57	**84.23**	53.15
ES	55.09	26.13	48.75	**81.84**

Table 7: NER results on the CoNLL test sets for EN-BERT. The row is the fine-tuning language, the column the evaluation language. There is a big gap between this model's zero-shot performance and M-BERT's, showing that the pre-training is helping in cross-lingual transfer.

C Some POS Results for EN-BERT

Fine-tuning \Eval	EN	DE	ES	IT
EN	**96.94**	38.31	50.38	46.07
DE	28.62	**92.63**	30.23	25.59
ES	28.78	46.15	**94.36**	71.50
IT	52.48	48.08	76.51	**96.41**

Table 8: POS accuracy on the UD test sets for a subset of European languages using EN-BERT. The row specifies a fine-tuning language, the column the evaluation language. There is a big gap between this model's zero-shot performance and M-BERT's, showing the pre-training is helping learn a useful cross-lingual representation for grammar.