CS388 Final Project: A Broader Evaluation of Multilingual BERT on Code-Switched Data

Kelsey Ball, EID:ktb477 & Tongzheng Ren, EID:tr24699

Abstract

Cross-lingual language models provide meaningful representations of text in many different languages, enabling NLP research and technologies to benefit a broader global community. Pre-trained mulitlingual masked language models like M-BERT [9] and XLM [12] have demonstrated particularly strong performance on cross-lingual understanding benchmarks. However, it remains unclear how well such models represent *code-switched* text, where multiple languages alternately surface in the same sentence or context. Recent work on code-switching uses cross-lingual language models like M-BERT as a baseline encoder [2]; here, we attempt to more broadly characterize M-BERT's capacity to represent code-switched data. Consistent with prior work, our results show that M-BERT can perform reasonably even without access to code-switched data. We also show that few-shot learning with limited code-switched data provides a significant performance increase, indicating a non-trivial domain mismatch between monolingual and code-switched settings. Finally, we make some initial attempts to understand when and how M-BERT can perform well as an encoder for code-switched data.

1 Introduction

Deep contextualized language models can provide powerful representations of natural language, boosting performance on a variety of downstream natural language processing tasks [15, 9]. These models are generally pre-trained on large amounts of unannotated text, then fine-tuned on smaller amounts of supervised data to handle specific tasks. Previous work has shown that such representations can encode syntactic and named-entity information, thus generalizing well even with limited annotated data [16, 20, 19]. However, prior work focuses on models trained on a single language, particularly English.

To mitigate this English-centered bias, research has increased on cross-lingual language models, whose aim is to provide powerful representations regardless of the language of the input sequence [9, 12, 6]. The multilingual BERT (M-BERT) model released by [9] trains a single language model on the Wikipedia corpora from 104 different languages using a shared wordpiece vocabulary, without explicit markers denoting the input language or any explicit mechanisms forcing translation-equivalent pairs to share similar representations. [21, 17] demonstrate the effectiveness of M-BERT on providing multilingual representations by showing that M-BERT has surprisingly good zero-shot cross-lingual transfer performance on POS-tagging and NER. [6] further shows that, when scaling the amount of training data, such zero-shot cross-lingual transfer capability further improves. Recently, [10] suggests that lexical overlap between languages plays a significant role for the cross-lingual transfer of M-BERT.

Nonetheless, these works mostly focus on transfer between monolingual corpora. Code-switching is a distinct but related phenomenon in which multilingual speakers alternate between two or more languages in the context of a single conversation. Code-switching is ubiquitous in multilingual societies [5], but still understudied in NLP research. Particularly little work has examined the representation power of cross-lingual language models on code-switched data. The main exception

Preprint. Under review.

is [17], which shows that, on Hindi-English POS-tagging data from [4], M-BERT fine-tuned on monolingual data can achieve performance comparable to the model fine-tuned on code-switched data, indicating that M-BERT can learn decent representations of code-switched data from monolingual corpora alone. Other recent work [2] cites M-BERT as a baseline model for downstream tasks on code-switched data. We aim to complement these works by investigating the representational power of M-BERT for code-switched data and the degree to which it makes sense as an approach for modeling code-switching.

In this paper, we aim to conduct a more thorough evaluation of M-BERT on code-switched data and provide some interpretations on when and how M-BERT can perform well on code-switched data. Specifically, we extend the work of [17] by extending some of their experiments to additional language pairs to examine how well their conjectures generalize to other code-switched language pairs. We further propose some hypotheses on how and to what extent M-BERT performs well on code-switched data and provide experiments to corroborate our hypotheses.

Our empirical results show that, unlike the observation in [17], M-BERT fine-tuned on monolingual data does not necessarily lead to comparable performance when fine-tuned on code-switched data. We conjecture that the difference is due to a domain mismatch between monolingual data and code-switched data, and provide some empirical evidence in support of this conjecture. We further show that this domain mismatch can be alleviated with small amounts of code-switched data. Finally, we provide an initial attempt to understand when and how M-BERT can perform well on code-switched data, by understanding the feature-space alignment between different languages using the nearest-neighbor accuracy metric proposed in [17]. We hope our work can help in understanding the representational power of cross-lingual language models, as well as improving techniques for code-switched language processing.

2 Related Work

Cross-Lingual Language Models Cross-lingual language models gained attention in recent NLP research for their strong performance on tasks which require cross-lingual understanding. Multilingual masked language models like M-BERT [9] and XLM [12] have pushed the state-of-the-art on cross-lingual understanding benchmarks like cross-lingual natural language inference [7], question answering [13] and named entity recognition [17, 21], by jointly training the models on monolingual Wikipedia corpora from different languages. Later, [6] further boosted the performance by scaling the amount of pre-training corpora. Several hypotheses have been proposed to understand the success of these models. [17] suggests that typological similarities can be helpful for cross-lingual understanding, while [6] suggests that lexical overlap between languages plays the most important role. Recently, [8] argues that the only requirements for cross-lingual transfer is the parameter sharing on the top layer of the multi-lingual encoder, which provides a different perspective on the importance of the model.

Code-Switched Language Processing Code-switching is a linguistic phenomenon which occurs when a multilingual speaker alternates between different languages in the context of single sentence of conversation. It is widely observed in informal web text, especially on social media platforms like Twitter and Facebook. Code-switched data brings a relatively new challenge to the NLP community, which traditionally has focused primarily on monolingual and multilingual settings; as such, the main difficulty for the research on code-switched data is the relative lack of (annotated) code-switched data. We refer the readers to [18] for a survey on code-switched language processing.

Recently, [1] proposed a new centralized benchmark, LinCE, for the evaluation of different models across 15 code-switched datasets and 4 language pairs. Additionally, the LinCE benchmark proposes new splits for existing code-switched datasets which are more balanced and carefully chosen. Similarly, [11] introduce another benchmark for code-switched evaluation called GLUECoS, which focuses on two code-switched language pairs, but introduces code-switched datasets addressing higher-level NLP tasks like inference and question-answering.

These benchmarks represent an important effort to unify research in code-switched language processing; we briefly mention why we evaluate against some but not all of the datasets they comprise. Because M-BERT is trained on Wikipedia corpora of many different languages, the majority of its training data is written in the standard script of each input language. Consequently, we expect M-BERT to be less capable of transferring to a transliterated target; indeed, Pires et al. [17]

demonstrate this empirically on transliterated, code-switched Hindi-English. In this work, we aim to evaluate M-BERT in more favorable settings that circumvent the challenge of transliteration. For this reason, instead of evaluating against either of the LinCE or GLUECoS benchmarks which include transliterated code-switching like Hindi-English and Nepali-English, we focus instead on evaluating code-switched language pairs whose surface forms are individually well-represented in M-BERT's training data: specifically, Spanish-English and Turkish-German.

3 Evaluation of M-BERT on Code-Switched Data

In this section, we provide a more thorough evaluation of M-BERT on different code-switched pairs of languages. Follow the setting of [17], we compare the performance of M-BERT on downstream tasks like POS-tagging after fine-tuning on monolingual data in order to see if M-BERT can have good empirical performance on code-switched data without observing any. However, unlike [17], we evaluate on additional language pairs to see if M-BERT can have good empirical performance uniformly across different language pairs, particularly those whose scripts M-BERT has seen during training. Moreover, when fine-tuning with monolingual data, we use different kinds of source corpora to investigate the impact of genre on the final performance.

English-Spanish Code-Switching We use part-of-speech (POS) tagging as our evaluation task. The code-switched data we use are from the LinCE benchmark [1], which contains 27893 sentences in the training set and 4298 sentences in the development set. For monolingual data, we use corpora from Universal Dependencies [14]: specifically, we use the Ancora corpus for Spanish data and the EWT and GUM corpora for English data. The results are shown in Table 1.

Table 1: POS-tagging results for English-Spanish Data.

Train	Development	F1	Accuracy
English EWT, Spanish Ancora	English-Spanish LinCE	85.76	87.17
English GUM, Spanish Ancora	English-Spanish LinCE	84.81	86.17
English-Spanish LinCE	English-Spanish LinCE	96.36	96.99

Our results show that fine-tuning on monolingual data can produce acceptable performance on code-switched data; however, fine-tuning on monolingual corpora alone is not as effective for code-switched English-Spanish when compared to the result from [17] on Hindi-English, which only has 4% accuracy gap (compared to $\sim 10\%$ here) between fine-tuning on monolingual vs. code-switched data.

Pires et al. posit that this gap comes not from limited representation power but from a domain mismatch in the genre of the training and evaluation corpora: *"it is likely that some of the remaining difference is due to domain mismatch"*. This hypothesis is plausible, given that the monolingual Hindi and English corpora consist of formal news text, while the code-switched evaluation text is informal web/social media text.

We examine the potential effect of domain mismatch between the training and evaluation corpora by finetuning M-BERT on two different English corpora: the GUM corpus, which consists of variety of genres including fiction and biographies, and the EWT, a web-text corpus that lies much closer in domain to the evaluation text. Despite being closer in genre, fine-tuning on EWT does not produce a significant performance increase ($\sim 1\%$ accuracy). This suggests that the performance gap is not dominated by the domain mismatch in the text genre, but rather a more structural domain mismatch between the monolingual and code-switching settings.

To further demonstrate this potential structural domain mismatch, we also evaluate the fine-tuned model on the monolingual development set. The results are shown in Table 2.

Here we can see fine-tuned model has outstanding performance on each monolingual development set, indicating that the fine-tuned model perfectly learned the structure of the monolingual data, and the performance drop is probably not due to under-fitting of the monolingual data, but rather a structural domain mismatch between monolingual and code-switched data.

Table 2: POS-tagging results for monolingual English and Spanish Data.

Development	F1	Accuracy
English EWT	95.87	96.46
Spanish Ancora	98.78	98.96

Turkish-German Code-Switching We extend our POS-tagging evaluation to code-switched Turkish-German, using data from the SAGT project[1]. There are 285 public available training sentences as well as 801 development sentences. We again use monolingual corpora from Universal Dependences: the GSD corpus for German data and the BOUN corpus for Turkish data. The results are shown in Table 3.

Table 3: POS-tagging results for German-Turkish Data.

Train	Development	F1	Accuracy
German GSD, Turkish BOUN	Turkish-German SAGT	74.82	78.48
Turkish-German SAGT	Turkish-German SAGT	85.98	89.19

These results again show that fine-tuning on monolingual data can have acceptable empirical performance, but there is still a significant performance gap when compared to fine-tuning on code-switched data.

Similarly, to rule out the potential issue of under-fitting on the monolingual data, we evaluate the fine-tuned model on the monolingual development sets. The results are shown in Table 4.

Table 4: POS-tagging results for German-Turkish Data.

Development	F1	Accuracy
German GSD	95.47	96.03
Turkish BOUN	86.87	90.61

Again, the fine-tuned model has much better performance on the monolingual development data than on code-switched data. We note that the performance on Turkish BOUN development set is not directly comparable to the performance on other monolingual data, given that the Turkish training set is much smaller (the Turkish BOUN training set has roughly half the number of sentences as the English and German training sets). Nonetheless, it's performance is still significantly better than in the code-switched .

In summary, our observation suggests that we cannot necessarily expect a POS-tagging system fine-tuned only on monolingual data to perform well on code-switched data; additional information is needed to make M-BERT have better performance on code-switched data.

Code-Switching vs. Concatenating Languages M-BERT is trained on the concatenation of Wikipedia corpora in 104 different languages. The accepted usage of M-BERT as a encoder for code-switched targets may lead to the following naïve interpretation of code-switching:

Is code-switching just a concatenation of different languages?

Scholarship on code-switching [5] identifies it as quite distinct from the sum of its component languages; however, the relative paucity of code-switched data (and relative abundance of monolingual data) encourages acceptance of this overly simple model of code-switching among practitioners.

To empirically refute this view of code-switching, we evaluate against a synthetic code-switched development set in which each sentence is the concatenation of two sentences from different languages. We test the performance on this synthetic code-switched development set after we fine-tune the model on monolingual data. The results are shown in Table 5.

We can see that the model fine-tuned on monolingual data works well on this synthetic development set; this is unsurprising, as the model likely handles each monolingual sentence separately. Meanwhile,

[1] https://github.com/UniversalDependencies/UD_Turkish_German-SAGT/tree/master

Table 5: POS-tagging results on naïve synthetic CS data.

Train	Development	F1	Accuracy
English EWT, Spanish Ancora	Synthetic English EWT + Spanish Ancora	97.17	97.49
German GSD, Turkish BOUN	Synthetic German GSD + Turkish BOUN	90.62	92.6

there remains a huge performance gap between real and synthetic code-switched data, indicating that real code-switched data is not just the simple combination of two individual languages. Rather, it is more likely that code-switching has a unique structure and domain, the learning of which requires information beyond monolingual examples from each constituent language.

Alleviate Domain Mismatch with Few-Shot Learning So far, we have shown that parts of the performance gap can be attributed to a domain mismatch between monolingual and code-switched data. Here we explore to what extent that domain mismatch can be alleviated with few-shot learning. Intuitively, given a good representation of some data, a model can adapt to a new domain with little amounts of data, an approach commonly referred to as few-shot learning. To gauge the size of this potential domain mismatch, we provide the model with a small amount of code-switched data after fine-tuning with a large amount of monolingual data. The results on English-Spanish LinCE are shown in Table 6.

Table 6: Few-Shot Learning on English-Spanish LinCE. We first fine-tune with English EWT and Spanish Ancora, then fine-tune with small amount of code-switched data.

% of Code-switched training data	F1	Accuracy
1	90.83	92.29
5	94.82	95.67
10	95.18	95.98
20	95.71	96.42
100	96.36	96.99

As we can see, the performance dramatically improves after fine-tuning with small amounts of the code-switched data and approaches the performance of fine-tuning with entire dataset while using only 20% of the training data. These results partially demonstrate that the performance gap is due to learnable domain mismatch between monolingual and code-switched data.

We also test on the Turkish-German SAGT, and the results are summarized in Table 7.

Table 7: Few-Shot Learning on Turkish-German SAGT. We first fine-tune with German GSD and Turkish BOUN, then fine-tune with small amount of code-switched data.

% of Code-switched training data	F1	Accuracy
1	74.83	78.53
5	75.66	79.45
10	75.81	79.65
20	75.84	79.68
50	78.71	82.35
100	85.98	89.19

Here we notice that the performance on Turkish-German SAGT is not as good as English-Spanish LinCE; however, this is likely due to the limited amount of Turkish-German SAGT data. (Recall that there are only 285 public available training sentences in Turkish-German SAGT, but 27893 sentences in the English-Spanish LinCE.) Thus, the data we provide for fine-tuning on code-switched data is probably insufficient for the model to learn the specific structure of code-switching. Nonetheless, fine-tuning with half of the code-switched data still produces a significant performance gain.

Therefore, if the code-switched data is available, we recommend fine-tuning with code-switched data after fine-tuning with large amounts of monolingual data; this approach will likely yield significant benefit when dealing with the code-switched inputs in real-world applications.

Adding Explicit Language Signal In Pires et al.'s [17] experiment on Hindi-English POS tagging, they observe comparable performance between fine-tuning on monolingual corpora vs. code-switched corpora; however, they evaluate against script-corrected Hindi, where romanized Hindi tokens have been back-transliterated into Devanagari. One hypothesis for the success of this experiment is that evaluating on script-corrected tokens provides an explicit language signal to M-BERT: if a token is in Roman script, it is known to be English; if it is written in Devanagari, then it is known to be Hindi. It is plausible that this extra signal helps the model with the downstream classification task; many previous approaches to code-switched POS tagging ([4], [3]) show that supplying gold language tags to the model can improve accuracy.

Here we examine this hypothesis for Spanish-English and Turkish-German, which use shared scripts in the code-switched setting. We inject an explicit language signal by appending a language tag to each token during training and evaluation. Here we still fine-tune the model on the monolingual data and evaluate on the code-switched data. The results are summarized in Table 8.

Table 8: POS-tagging results with Explicit Language Signal.

Train	Development	F1	Accuracy
English EWT, Spanish Ancora	English-Spanish LinCE	86.15	87.57
German GSD, Turkish BOUN	Turkish-German SAGT	75.06	79.02

Surprisingly, we do not see any significant performance gain with an added explicit language signal ($< 1\%$). We remark that this is potentially a methodological issue; it's plausible that the appended language tags are not tokenized consistently or effectively by the model, whereas in [3] the LSTM-based architecture allows simply appending an additional binary feature to each token representation. As such, this hypothesis needs further investigation.

4 How and When Can M-BERT Encode Code-Switched Data?

So far we have shown that the good performance of M-BERT on code-switched data is not universal across language pairs and that the performance gap can be partially attributed to a structural domain mismatch between monolingual and code-switched data. Finally, we want to investigate how and when M-BERT works well for code-switched data.

Nearest Neighbour Accuracy from [17] One hypothesis is that, in order for M-BERT to work well on a given language pair, the feature space of M-BERT should align well for the two languages. To evaluate the degree of multilingual learning in M-BERT, [17] compute a measure called "nearest neighbor accuracy" between two languages. For the l-th layer of M-BERT, we can compute a sentence embedding $v^{(l)}$ by averaging the feature representation of all tokens except `[CLS]` and `[SEP]`. Then, for pairs of translation-equivalent sentences, we compute the following vector:

$$\bar{v}^{(l)}_{\text{LANG}_1 \to \text{LANG}_1} = \frac{1}{M} \sum_{i \in [M]} \left(v^{(l)}_{\text{LANG}_1, i} - v^{(l)}_{\text{LANG}_2, i} \right),$$

where $v^{(l)}_{\text{LANG}_1, i}$ and $v^{(l)}_{\text{LANG}_2, i}$ are the embeddings of the translation-equivalent sentence pair i at layer l. $\bar{v}^{(l)}_{\text{LANG}_1 \to \text{LANG}_1}$ can thus be interpreted as a "translation" vector from LANG_1 to LANG_2. We then use $\bar{v}^{(l)}_{\text{LANG}_1 \to \text{LANG}_1}$ to translate the sentence from LANG_1 to LANG_2, and measure the fraction of times the closest sentence (in terms of ℓ_2 distance) is the exact translation from LANG_1.

Here we extend this experiment to a code-switched setting by evaluating translation pairs which include code-switching. Our goal is to see if M-BERT learns a semantically aligned feature space with respect to code-switching. We use data from the ACL 2021 Shared Task on Machine Translation in Code-Switching Settings[2], which includes aligned sentence pairs in English and code-switched Hindi-English, or "Hinglish".

[2]https://www.aclweb.org/portal/content/call-shared-task-pariticipation-machine-translation-code-switching-environments

Due to a lack of more aligned, code-switched data, we are unable to extend our evaluations to other code-switched language pairs; however, we compute nearest-neighbor accuracy for additional monolingual language pairs which were not evaluated in Pires et al.: English to German (`en->de`), English to Hindi (`en->hi`), English to Turkish (`en->tr`), and English to Spanish (`en->es`). We use the WMT16 newstest2015 parallel corpus for English-German translation, PMIndia parallel corpus for English-Hindi translation, a parallel news corpus called Bianet for English Turkish translation, and the WMT13 Common Crawl corpus for English Spanish translation. As the WMT16 newstest2015 parallel corpus contained the fewest (2169) samples, we use 2169 samples for all of the translation tasks. (Note that a random baseline for this task produces $< 1\%$ accuracy.) The results are shown in Figure 1.

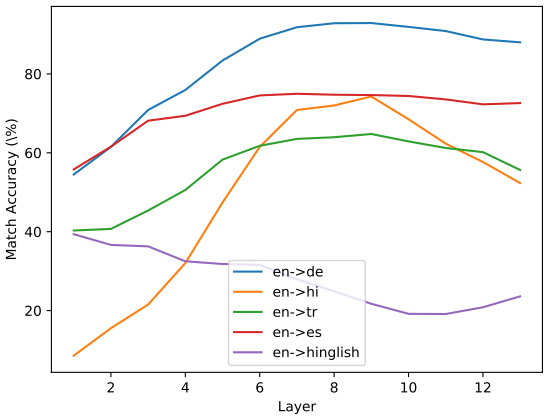

Figure 1: Nearest-neighbor accuracy of different language pairs.

We can see the nearest-neighbor accuracy for most of the language pairs first increases with the depth then slowly decreases, matching the intuition that the higher layers contain more semantic information, while the lower layers contain more language-specific information. We posit that alignment in a given languages pair predicts reasonable performance on their code-switched form; this could be suggested by our POS-tagging results on English-Spanish and those of [17] on English-Hindi, combined with the reasonable nearest-neighbor accuracy we see for both language pairs. However, due limited parallel data, we currently don't have more evidence support this argument. We also note that the nearest-neighbor accuracy for English and Hinglish mostly decreases with model depth, which is not consistent with other language pairs. This indicates that M-BERT does not necessarily align the feature space between monolingual and code-switched data, despite aligning the constituent languages of a given pair. One interesting observation is that the nearest-neighbour accuracy peaks at layer 8 for most languages. A natural question is: can we derive better performance on cross-lingual tasks using the representation from the 8-th layer? We leave this as future work.

5 Conclusion

We demonstrate that, although M-BERT can perform reasonably well on code-switched data using only monolingual resources, this is not universal across language pairs. We attribute the performance gap to a structural domain mismatch between monolingual and code-switched data, which can be alleviated with small amounts of code-switched data. We also try to understand the mechanism behind M-BERT's representation power on code-switched data, but due to resource constraints, we mostly conjecture that good alignment in the feature space of two languages predicts better representations of their code-switched form. Future work can include collecting more code-switched data for a larger-scale evaluation of code-switched data, developing alternative strategies for alleviating the domain shift without access to code-switched data, and better characterizing the ability of cross-lingual models to represent code-switched data.

References

[1] Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, 2020.

[2] Gustavo Aguilar and Thamar Solorio. From english to code-switching: Transfer learning with strong morphological clues. *CoRR*, abs/1909.05158, 2019.

[3] Kelsey Ball and Dan Garrette. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3084–3089, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[4] Irshad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava, and Dipti Misra Sharma. Universal dependency parsing for hindi-english code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, 2018.

[5] B.E. Bullock and A.J. Toribio. *The Cambridge Handbook of Linguistic Code-switching*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2009.

[6] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[7] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, 2018.

[8] Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

[10] K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2019.

[11] Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online, July 2020. Association for Computational Linguistics.

[12] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, 2020.

[14] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[15] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[16] Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, 2018.

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

[18] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*, 2019.

[19] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

[20] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2018.

[21] Shijie Wu and Mark Dredze. 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, 2019.