Can Multilingual Language Models Transfer to an Unseen Dialect? A Case Study on North African *Arabizi*

Benjamin Muller **Benoit Sagot** **Djamé Seddah**
INRIA, Paris, France
firstname.lastname@inria.fr

Abstract

Building natural language processing systems for non standardized and low resource languages is a difficult challenge. The recent success of large-scale multilingual pretrained language models provides new modeling tools to tackle this. In this work, we study the ability of multilingual language models to process an unseen dialect. We take user generated North African Arabic as our case study, a resource-poor dialectal variety of Arabic with frequent code-mixing with French and written in Latin script. Focusing on two tasks, part-of-speech tagging and dependency parsing, we show in several scenarios that multilingual language models are able to transfer to such an unseen dialect, specifically in two extreme cases: (i) across scripts, using Modern Standard Arabic as a source language, and (ii) from a distantly related language, unseen during pretraining, namely Maltese. Our results constitute the first successful transfer experiments on this dialect, paving the way for the building of an NLP ecosystem for *creole*-like, resource-scarce languages.

1 Introduction

Accurately modeling low resource and non-standardized languages exhibiting a high degree of variation is extremely challenging. Recent releases of multilingual language models trained on large corpora (Devlin et al., 2019; Lample and Conneau, 2019) provide an interesting opportunity to address this challenge in new ways. We frame our work as a *cross-lingual transfer learning* analysis: we study the capacity of a system trained as a language model on a source set of languages to transfer to a target language and task. More precisely, we investigate the ability of multilingual language models to process a language that is absent from their pre-training set. For brevity, we simply refer to such languages as *unseen*.

Our work focuses on the multilingual version of BERT (mBERT) (Devlin et al., 2019). The cross-lingual modeling ability of mBERT has been recently studied by Pires et al. (2019), who show that cross-lingual transfer is very efficient between pretrained languages. In our work, we address a different and more challenging question: can mBERT transfer to an unseen and non-standardized dialect? We take North-African Arabizi, hereafter *Narabizi*, as our case study. We define *Narabizi* as the romanization[1] of the Arabic dialect spoken in Algeria, found ubiquitously on social media. It is a non-standardized dialect with no standard writing system and shows a high degree of code-mixing with French (Amazouz et al., 2019). This makes *Narabizi* highly variable across users and therefore very challenging for Natural Language Processing.

For our experiments, we use the *Narabizi* raw corpus and treebank recently released by Seddah et al. (2020) and focus on two tasks, namely part-of-speech (POS) tagging and dependency parsing. After a detailed cross-lingual performance analysis, our results show that multilingual models are able to transfer to unseen, highly variable data. More precisely, we make the following contributions:

- We push the zero-shot cross-lingual abilities of mBERT to the extreme and show that it can transfer to unseen *Narabizi* in POS tagging and parsing, even when the source is another unseen and related language such as Maltese
- By running comparison across source languages and diverse BERT models, we demonstrate that mBERT is using its multilingual representations to process *Narabizi*.
- We show the positive impact of unsupervised fine-tuning on cross-lingual transfer and demonstrate its ability to make transfer possible, even across scripts, in a scenario where the target language is not in the pre-training corpora.

[1] i.e. written in Latin script.

2 Related Work

Word embedding & Cross lingual Transfer Recently, cross lingual transfer has benefited from multilingual language models. We refer to (Lample and Conneau, 2019; Eisenschlos et al., 2019; Vania et al., 2019; Wu et al., 2019; Conneau et al., 2019; Wu and Dredze, 2019) who demonstrate the efficiency of language models in zero-shot transfer settings for a variety of tasks. In this regard, Pires et al. (2019) analyze in detail the zero-shot transfer ability of mBERT on sequence labeling. Wang et al. (2019) suggest that cross-lingual transfer of multilingual models rely on *structural properties* of languages. Both studies focus on transfer between languages that are part of the pretraining corpora. In our work, we study the ability of mBERT to transfer to an unseen language.

Code-Switching is a hard challenge for NLP as shown in the myriad of works that have tackled this phenomenon for more than 10 years, see for example (Solorio and Liu, 2008; Vyas et al., 2014; Çetinoğlu and Çöltekin, 2016; Lynn and Scannell, 2019). Ball and Garrette (2018) and Pires et al. (2019) analyzed the performance of neural models for sequence labeling showing that those approaches can cope with such a complexity. In our work, we face both code-switched and highly variable data.

Unsupervised Adaptation of Language Models Han and Eisenstein (2019) show that fine-tuning BERT in an unsupervised way using its masked language objective brings significant improvement to downstream sequence labeling tasks for out-of-domain Old English. Studying the specific case of English-Spanish code-mixing, Gonen and Goldberg (2018) show how to adapt bilingual language models to code-mixed data. In our work, we focus on unsupervised adaptation and analyze its impact on the even more challenging case of *Narabizi*.

3 *Narabizi*

Arabic varieties are often classified into three categories (Habash, 2010): (i) Classical Arabic, as found in the *Qur'an* and related canonical texts, (ii) Modern Standard Arabic (MSA), the official language of the vast majority of Arabic speaking countries and (iii) Dialectal Arabic. This work focuses on North-African dialectal Arabic in its Algerian form, understood and spoken by more than 40 million people in the Maghreb (Sayahi, 2014). In its written form, it is mostly found online and in Latin script. For simplicity we refer to this North-African Arabic dialect as North-African Arabizi (Farrag, 2012) or *Narabizi*, illustrated here:

source: `Mrhba, Ana 3rbi mn dzaye`
translation: *"Hey, I'm Arab from Algeria"*

Like other written languages found on social media and even more importantly as it is not standardized,[2] *Narabizi* shows a high degree of variability across writers. As part of its variability, *Narabizi* frequently involves code-switching with French.

Moreover, *Narabizi* does not belong to the pretraining corpora of mBERT. For this reason, we take *Narabizi* as our case study to analyze the ability of mBERT to handle an unseen, highly variable and code-mixed dialect.

Data The data we use comes from two main sources. The first one, described by Cotterell et al. (2014), is a collection of 9000 raw Algerian romanized Arabic sentences, a sample of which has been annotated with Universal Dependency trees (McDonald et al., 2013) and word-level language identification[3] by Seddah et al. (2020) totalling 1,434 (1172/146/178) annotated sentences. Our second source, also released by Seddah et al. (2020), is a collection of 49,546 raw *Narabizi* sentences.

Baselines To grasp the complexity of *Narabizi*, we run some preliminary experiments. We take Qi et al. (2019)'s tagger and parser as our strong baselines (named StanfordNLP,[4]) and as our bottom lines the majority class predictor for POS tagging and the *left predictor*[5] for dependency parsing. Competitive taggers perform on datasets of similar size above 90%. StanfordNLP only reaches 84.20% on our data for POS tagging and 52.84% for parsing, as measured by the unlabeled attachment score (UAS; cf. Table 1).

4 Model

mBERT is a Transformer (Vaswani et al., 2017) trained as a joint masked-language and a next sentence prediction model on sub-word level tokenized sentences. More details can be found in (Devlin et al., 2019). We use the multilingual cased version

[2] I.e. no writing rules are officially defined.

[3] *Narabizi* and French prevalence in the train set (% token): *Narabizi* 64.64%, French 33.84%, then MSA, English & Spanish.

[4] Ranked top 3 (after correction) in POS tagging and parsing at the 2018 UD shared task, trained using French fastText vectors (Mikolov et al., 2018).

[5] Whereby each word is attached to its immediate left neighbor.

of BERT. mBERT was trained on the concatenation of the Wikipedia corpora for 104 languages.

POS tagging and dependency parsing with mBERT Following Devlin et al. (2019), we turn mBERT into a POS tagger by appending a softmax on top of its last layer. For parsing, we append the biaffine graph parser layers described by Dozat and Manning (2016). In both cases, we fine-tune the overall model by backpropagating only through the first sub-word token of each word. We call these architectures mBERT+POS and mBERT+PARSE.

Unsupervised Adaptation We call *unsupervised adaptation* the process of fine-tuning mBERT in an unsupervised manner using its Masked-Language Model (MLM) objective trained on raw sentences. We refer to mBERT fine-tuned on raw data as mBERT+MLM. We define as mBERT+MLM+TASK, with TASK referring to POS, resp. PARSE, to point to mBERT+MLM fine-tuned as a POS tagger, resp. parser.

5 Experiments

Our goal is to measure how well mBERT makes use of its multilingual pre-training on an unseen dialect. We defined a *source* language as a language on which a POS tagger or a parser are trained, and will report the performance of the resulting models when applied to *Narabizi* data.

5.1 Source Languages

We study transfer along two independent directions. The first one is the relatedness of the source language to *Narabizi*. The more different they are, the worse we expect the transfer to be. Our second direction distinguishes between source languages included in mBERT pre-training corpora and those that are not. We expect the transfer to be better when the source language is included in the pre-training corpora. To cover the full scope of cases, we pick Modern Standard Arabic, French, English and Vietnamese.

As recalled in (Habash, 2010; Čéplö et al., 2016), Maltese is related to the Arabic continuum of languages. It is standardized and written in an extended Latin script. This makes Maltese a promising candidate for transferring to *Narabizi*.

We also use French in order to study the impact of code-mixing. In addition, we experiment with English as another European language written in Latin script, but which is not code-mixed with *Narabizi*. Finally, we use Vietnamese as the most unrelated language to test the cross-lingual power of the model in the most extreme case. We refer the reader to the Appendix (Table 3) for an overview of the source languages. We sample the training datasets to have 1,200 sentences for each source language.[6] Additionally, we report results in the standard supervised setting in which we fine-tune mBERT+TASK on *Narabizi* and evaluate on *Narabizi*. This provides us with an upper bound on how we can expect mBERT to perform on such a language.

5.2 Optimisation

For supervised fine-tuning, we use the same range of hyper-parameters as Devlin et al. (2019). For unsupervised adaptation, we run preliminary experiments to measure the impact of the raw corpus among the 49,000 sentences *Narabizi* corpus, and *Narabizi* mixed with a sub-sample extracted from mBERT pre-training corpora. As reported by Gonen and Goldberg (2018), we found that, if carefully optimized, fine-tuning mBERT with its masked language objective directly on the target data leads to the best models.[7]

6 Results and Discussion

We present our results in Table 1. We report the accuracy for POS tagging and the Unlabeled Attachment Score (UAS) for parsing.[8] Any performance above the bottom line demonstrates that transfer is happening from the pre-training or fine-tuning stages to process *Narabizi*. For both POS tagging and parsing, mBERT+TASK performs outperforms the baselines by a large margin when the source is Maltese, French and English. This shows that mBERT is able to transfer to *Narabizi* even without having been trained on any *Narabizi* tokens at any stage of the training process.

We report an average boost with mBERT+MLM+TASK of +10.15 points in POS tagging and +4.10 in UAS across all source languages when compared with mBERT+TASK. This means that the unsupervised adaptation on 49,546 raw *Narabizi* sentences is efficient even on such an out-of-domain language. In all these settings, StanfordNLP, the very strong neural baseline, designed specifically for POS tagging and parsing, is outperformed by mBERT (it only reaches 31.90 and 33.74 (resp. 15.53) for parsing

[6] We pick the first 1,200 training sentences, more information on the datasets used is given in Appendix table 4.

[7] cf. Appendix § A.3 for details on hyper-parameters.

[8] For Labeled Attachment Score (LAS) cf. Appendix § A.2.

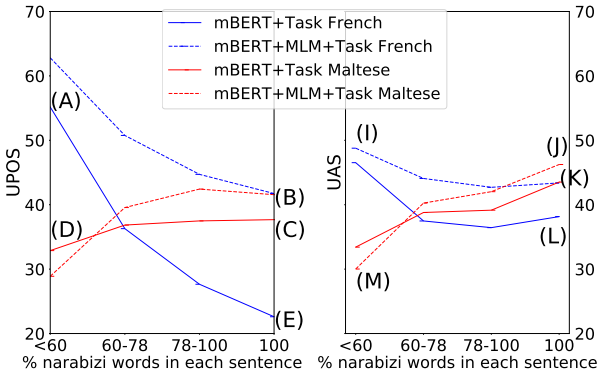

Figure 1: Performance with regard to code-mixing rate, reported on *Narabizi* train set to have enough data per bucket. (5 seeds). (X) markers commented in § 6.2. NB: no *Narabizi* annotated training data seen during fine-tuning.

	mBERT+TASK		mBERT+MLM+TASK	
Source	POS	UAS	POS	UAS
Maltese	**35.13**	**40.04**	38.94	42.32
French	33.12	38.54	**47.32**	**43.77**
English	30.67	32.40	44.59	38.83
MSA	16.55	28.23	28.08	34.34
Viet.	16.92	13.98	23.21	14.44
Narabizi	81.60	66.84	82.61	67.12
	Baselines			
	StanfordNLP		Bottom lines	
Narabizi	84.20	52.84	20.49	18.71
French	27.00	33.74	-	-

Table 1: Cross-Lingual performance on the *Narabizi* test set. 5 seeds averaged. Baselines described in § 3

.

in UAS (resp. in LAS) on *Narabizi* when trained on French)

6.1 Cross-script Transfer

In zero-shot settings, cross-script transfer does not perform above the bottom lines when the source is MSA in POS tagging. Our hypothesis is that such transfer requires the target language to be in the pre-training corpora as reported by Pires et al. (2019) in the case of Urdu and Hindi. Nevertheless, to our surprise, we observe an impressive +13 boost in tagging and +6.11 in parsing performance after unsupervised adaptation when the source is MSA, outperforming the baselines. This means that in the case of MSA, cross-script transfer happens when the target language is seen during unsupervised adaptation, and 49k sentences are enough to lead to such a transfer. Moreover, cross-script transfer is better with MSA than Vietnamese, suggesting that the multilingual model is making use of the proximity of MSA and *Narabizi*.

6.2 Impact of code-mixing

We hypothesize that the high level of transfer when the source is French is due to the high code-mixing proportion of *Narabizi*. To test our hypothesis, we present in Figure 1 the performance of the model with respect to the code-mixing ratio. We split the dataset into four buckets of around 25% of the full dataset, according to the ratio of native *Narabizi* tokens in each sentence (between less than 60% to stricly 100%) as opposed to French tokens. We compare French and Maltese as source languages. We confirm our intuition that code-mixing explains the good performance of the model trained on French. Indeed, on sentences that have 100% *Narabizi* tokens, mBERT+TASK trained on French performs poorly (cf. fig. 1 (E) for POS and (L) for parsing). On the other side, for sentences that include at least 40% of French tokens, scores reach 54% (cf. (A)) for POS tagging and 47% for parsing (cf. (I)). Moreover, for French, mBERT+MLM+TASK leads to an impressive 21.2% error reduction compared to mBERT+TASK for POS tagging (33.12 vs. 47.32) and an 8.5% error reduction for parsing (cf. Table 1). We observe in Fig. 1 (cf. (B) and (K)) that this improvement mostly comes from a better accuracy on *Narabizi* tokens. Interestingly, we observe that unsupervised fine-tuning leads to the closing of the gap between the performance of the models tuned on French and Maltese on native *Narabizi* tokens (+15: (B)-(E) vs. +2.4: (B)-(C) for POS tagging and +5.6: (K)-(L) vs. +2.2: (J)-(K) for parsing). This demonstrates the capacity of unsupervised fine-tuning to close lexical mismatch between distant languages such as native *Narabizi* and French.

6.3 Transfer between unseen languages

Surprisingly, mBERT+TASK tuned on Maltese does not perform poorly. It leads to the best performance for mBERT+TASK for both POS tagging and parsing. It outperforms StanfordNLP in the zero-shot scenario by 5 points in POS tagging and 6 points in parsing. As seen in Figure 1 (C) and (J), it performs the best on native *Narabizi* sentences (with no code-mixing). This result is surprising as Maltese is absent from the pre-training corpora. It shows that mBERT is able to capture *structural properties* shared by related languages even if they are absent from the pre-training corpora, thereby extending the observations described by Wang et al. (2019).

Is the multilingualism of mBERT at play? Finally, we want to show that the ability of mBERT to achieve cross-lingual transfer is related to the 104 languages it is pre-trained on, rather than because a pre-trained Transformer is an inherently good POS tagger or parser. To do so, we compare mBERT with three other models: Roberta, the optimized English version of BERT (Liu et al., 2019), CamemBERT (C.BERT) the French version of BERT (Martin et al., 2019), and a randomly initialized mBERT-like Transformer (Rand.). We focus our analysis on French and Maltese. mBERT is the model that leads to the most successful transfer in both cases and for both tasks, by a very large margin in the case of Maltese. This shows that pre-training on such a diversity of languages is at the core of the transfer to *Narabizi*.

	mBERT		RoBERTA		C.BERT		Rand.	
	POS	UAS	POS	UAS	POS	UAS	POS	UAS
fr	**33.12**	**38.54**	29.75	27.41	31.77	32.55	30.29	25.30
mt	**35.13**	**40.04**	25.45	17.27	31.62	34.65	19.81	19.04

Table 2: Zero-shot transfer from French (fr) and Maltese (mt) to *Narabizi*. 5 seeds averaged.

7 Conclusion

Our work on *Narabizi* reveals novel properties of multilingual language models. We have shown that their transfer ability can be pushed to the extreme of unseen target languages, even when transferring across script and from another unseen language.

Acknowledgments

We want to thanks Yanai Elazar, Ganesh Jawahar and Louis Martin for proofreading and insightful comments. This work was partly funded by two French National funded projects granted to Inria and other partners by the Agence Nationale de la Recherche, namely projects PARSITI (ANR-16-CE33-0021) and SoSweet (ANR-15-CE38-0011), as well as by the second author's chair in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. This project also received support from the French Ministry of Industry and Ministry of Foreign Affairs via the PHC Maimonide France-Israel cooperation programme.

References

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2019. Addressing code-switching in french/algerian arabic speech. In *Interspeech 2017*, pages 62–66.

Kelsey Ball and Dan Garrette. 2018. 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. Association for Computational Linguistics.

Slavomír Čéplö, Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, and Petr Zemánek. 2016. Mutual intelligibility of spoken maltese, libyan arabic, and tunisian arabic functionally tested: A pilot study. *Folia Linguistica*, 50(2):583–628.

Özlem Çetinoğlu and Çağrı Çöltekin. 2016. Part of speech annotation of a Turkish-German code-switching corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 120–130, Berlin, Germany. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An algerian arabic-french code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*.

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.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. 2019. Multifit: Efficient multi-lingual language model fine-tuning. *arXiv preprint arXiv:1909.04761*.

Mona Farrag. 2012. Arabizi: a writing variety worth learning? an exploratory study of the views of foreign learners of arabic on arabizi. Master's thesis, School of Humanities and Social Sciences, American University in Cairo, Cairo, Egypt.

Hila Gonen and Yoav Goldberg. 2018. Language modeling for code-switching: Evaluation, integration of monolingual data, and discriminative training. *arXiv preprint arXiv:1810.11895*.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan and Claypool.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *arXiv preprint arXiv:1904.02817*.

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

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*.

Teresa Lynn and Kevin Scannell. 2019. Code-switching in irish tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology Workshop*, pages 32–40, Dublin, Ireland. European Association for Machine Translation.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.

Sayahi. 2014. *The languages of the Maghreb. In Diglossia and Language Contact: Language Variation and Change in North Africa*. Cambridge: Cambridge University Press.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content north-african arabizi treebank :tackling hell. In *58th Annual Meeting of the Association for Computational Linguistics (ACL), Seattle, USA*.

Thamar Solorio and Yang Liu. 2008. Part-of-Speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii. Association for Computational Linguistics.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar. Association for Computational Linguistics.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

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.

A Appendix

A.1 Source Languages

Language	script	relatedness	$\in \Omega_{\text{mBERT}}$
Narabizi	Latin	-	no
French	Latin	code-mixed	yes
English	Latin	none	yes
Maltese	Latin	shared root	no
MS Arabic	Arabic	shared root	yes
Vietnamese	Latin	none	yes

Table 3: Source language in regard to *Narabizi* based on languages relatedness and inclusion in model pre-training corpora
NB: $\in \Omega_{\text{mBERT}}$ for languages included in the 104 pre-training languages of mBERT

Language	dataset
French	fr_gsd
MS Arabic	ar_padt
English	en_ewt
Maltese	mt_mudt
Vietnamese	vi_vtb

Table 4: Universal Dependencies (Nivre et al., 2016) Datasets used for cross-lingual experiments
(using the first 1,200 sentences when the treebank exceeds 1,200 sentences)

A.2 LAS scores

For parsing, we focus our cross-lingual analysis reporting for Unlabeled Attachment Score (UAS). We do so because we noticed significant annotation incompatibilities between the *Narabizi* treebank label scheme and the source treebanks labels, namely because of different annotation choices allowed by the Universal Dependencies framework. For the sake of completeness, we report the Labeled Attachment Score (LAS) in table 5.

Source	mBERT+PARSE	mBERT+MLM+PARSE
French	16.86	24.01
English	13.47	19.80
Maltese	17.81	18.10
MS Arabic	9.27	15.78
Vietnamese	3.60	4.75
Narabizi	56.60	57.94
	StanfordNLP	
Narabizi	32.5	-
French	15.53	-

Table 5: Parsing LAS scores in Cross-Lingual performance on *Narabizi* test set. Averaged over 5 seeds.

A.3 Hyper-parameters supervised and unsupervised fine-tuning

We list here all the hyper-parameters used for fine-tuning in a supervised way on POS tagging and in an unsupervised way on raw *Narabizi* data (cf. Table 6 and 7). For the supervised setting, we run a grid search on all the combination of hyper-parameters and select the best model on the validation set of the source language for both POS tagging and parsing.

parameter	value
batch size	{32,16}
learning rate	{1e-5,5e-5,1e-4}
optimizer	Adam
epochs (best of)	30

Table 6: Supervised fine-tuning hyper-parameters.

parameter	value
batch size	64
learning rate	5e-5
optimizer	Adam
warmup	linear
warmup steps	10% total
epochs (best of)	10

Table 7: Unsupervised fine-tuning hyper-parameters

A.4 Buckets : Detailed data and scores

Proportion Arabizi % of word in sentence	¡60	60-78	78-100	=100
train set number sents	322	286	283	276

Table 8: Code-mixed Buckets