Monolingual Pre-trained Language Models for Tigrinya

Fitsum Gaim, Wonsuk Yang, Jong C. Park
School of Computing
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Daejeon, Republic of Korea
{fgaim, dirrick0511, park@nlp.kaist.ac.kr}

Abstract

Pre-trained language models (PLMs) are driving much of the recent progress in natural language processing. However, due to the resource-intensive nature of the models, under-represented languages without sizable curated data have not seen significant progress. Multilingual PLMs have been introduced with the potential to generalize across many languages, but their performance trails compared to their monolingual counterparts and depends on the characteristics of the target language. In the case of the Tigrinya language, recent studies report a sub-optimal performance when applying the current multilingual models. This may be due to its orthography and unique linguistic characteristics, especially when compared to the Indo-European and other typologically distant languages that were used to train the models. In this work, we pre-train three monolingual PLMs for Tigrinya on a newly compiled corpus, and we compare the models with their multilingual counterparts on two downstream tasks, part-of-speech tagging and sentiment analysis, achieving significantly better results and establishing the state-of-the-art. We make the data and trained models publicly available.[1]

1 Introduction

Multilingual pre-trained language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are trained on large corpora of many languages and expected to learn useful representations that can potentially generalize to languages even unseen during training. While these models present a promising avenue for low-resource languages, the majority of the +7000 languages in the world are not covered during training, and typological diversity and other linguistic differences pose challenges, making their performance subpar compared to monolingual models (Muller et al., 2021; Rust et al., 2021; Ebrahimi et al., 2021). For example, multilingual PLMs show poor performance for Tigrinya even compared to other languages that were also not included during training (Wang et al., 2020). This fact may indicate that the multilingual models are not covering the characteristics that differentiate Tigrinya from the other well-known languages included in the models. Tigrinya [iso: tir] is one of the severely low-resourced languages of East Africa with around nine million speakers,[2] which lacks standard datasets and benchmarks for basic tasks such as language modeling. Tigrinya uses the Ge'ez script as a writing system, which is shared with Amharic and Tigre. In addition to being severely low-resourced, Tigrinya has a highly inflectional agglutinative morphology, presenting its own challenge to language modeling, particularly in contrast to the well-known and resource rich languages such as English. Therefore, monolingual language modeling is necessary for Tigrinya until much powerful multilingual methods are proposed. To this end, we compile a new corpus, build three monolingual PLMs for Tigrinya, and compare them against their multilingual counterparts on two downstream tasks – sentiment analysis and part-of-speech tagging.

[1] https://github.com/fgaim/tigrinya-plms
[2] https://www.ethnologue.com/language/tir, accessed on 2021-10-11.

Model	POS	Sentiment	#Params
mBERT	37.21	57.20	167M
mBERT (Transliterated)	91.24	81.48	»
XLM-R	<u>91.87</u>	<u>82.17</u>	278M
XLM-R (Transliterated)	90.92	81.81	»
TiELECTRA (Ours)	93.12	82.29	14M
TiBERT (Ours)	92.89	82.06	110M
TiRoBERTa (Ours)	**95.49**	**84.76**	125M

Table 1: Results of POS (Accuracy) and Sentiment Analysis (F1 score) of the multilingual and monolingual models. The underlined scores are the highest among multilingual models, and the bold-face scores are the best overall.

2 Pre-training Data and Models

Data: We could not find publicly available and sufficiently large data for Tigrinya language modeling, therefore, we compiled a new corpus for the task. The data was crawled from popular Tigrinya news sites, blogs, and books, with the majority of the text, ∼75%, coming from more than 2150 issues of an Eritrean periodical newspaper, *Haddas Ertra*,[3] that were published over a period of eight years, 2014-2021. In addition to news, the data contains articles on diverse domains such as education, health, science & technology, and fiction & non-fiction prose; hence we believe it is representative of the contemporary Tigrinya language. As part of pre-processing, we filter out only text written in Ge'ez script, remove text in legacy non-standard encoding systems, normalize characters, and remove foreign words. The final corpus contains over 40 million tokens, ∼2 million sentences, and ∼1 million vocabulary with 36% hapax legomena. We split the data into training and validation parts by the ratio of 98%-2%, respectively.

Models: We pre-train three transformer-based (Vaswani et al., 2017) language models on the corpus: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020). We chose BERT and RoBERTa for a fair comparison with mBERT and XLM-R, respectively, and account for performance impact due to the differences in model architecture or training objectives. ELECTRA-small, with only 14M parameters, is selected to explore the performance of comparatively small models with a sample-efficient training objective. All the other monolingual and multilingual models have over 100M parameters, as shown in Table 1. Following common experimental setups, we set the vocabulary size of RoBERTa to 50K, while BERT and ELECTRA both have 30K. All three models are trained for 40 epochs, with a maximum input sequence length of 512, and a mini-batch size of 128. Training is conducted using the Flax framework (Heek et al., 2020) and the Transformers library (Wolf et al., 2020) on a single Cloud TPU v3.8.

3 Evaluation on Downstream Tasks

We evaluate the performance of the pre-trained monolingual and multilingual models on two downstream tasks using existing data sets. All models are fine-tuned for 3 epochs using a maximum input sequence length of 128 in both tasks, while other hyper-parameters are task-specific. Our models, with much fewer parameters, are able to outperform their multilingual counterparts by significant margins.

POS: Tedla et al. (2016) prepared a Tigrinya POS dataset with a 20-class tagset containing 4.6K manually annotated sentences. We shuffle and split the data into train and test parts with a ratio of 8:2, respectively, and use that to compare the multilingual and monolingual models. We use a mini-batch size of 8 and a learning-rate of *5e-5*. The TiRoBERTa model outperforms all others, showing a gain of 3.64 points accuracy ahead of the strongest multilingual model,

[3]Ministry of Information, Eritrea, `www.shabait.com`

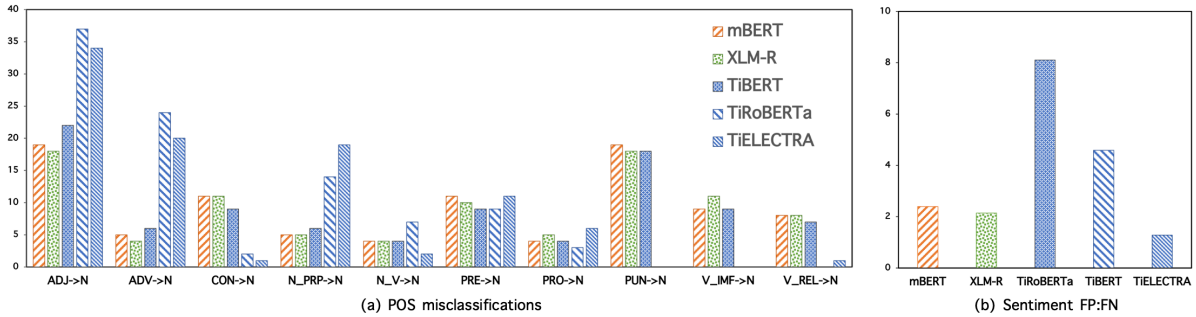

Figure 1: (a) Top ten most frequent mistakes in the POS task. 'X->Y' indicates a misclassification where the gold label X is predicted as label Y. (b) The ratio of False Positives to False Negatives in the sentiment analysis task.

XLM-R. For a fair comparison with Tedla et al. (2016), we perform a stratified 10-fold evaluation of TiRoBERTa and obtain 95.08 in accuracy, ∼4 points ahead of the previous best. As shown in Figure 1(a), the top ten most frequent misclassifications appear when the models predict the majority class, *noun*, with the *adjective* and *adverb* classes being the most confused ones. The multilingual models tend to make diverse errors compared to the monolingual ones.

Sentiment Analysis: Tela et al. (2020) developed a Tigrinya sentiment analysis dataset from YouTube comments, comprising a training set of ∼50K automatically labeled samples and a test set of 4K manually annotated examples. The data has two categories, positive and negative, with a balanced distribution across the dataset. We use a mini-batch size of 32 and a learning-rate of *2e-5* during fine-tuning. Once again TiRoBERTa performs the best, showing a gain of 2.59 points over the strongest multilingual model, XLM-R, and also improves the previous best 83.29 (Tela et al., 2020) by 1.47 points, establishing the state-of-the-art. Figure 1(b) shows the ratio of false positives and false negatives for all models, with TiRoBERTa showing the highest bias towards positive sentiment although it has the best overall accuracy.

Our results show that the smallest monolingual model, TiELECTRA, consistently outperforms the multilingual models in both downstream tasks, even though it has orders of magnitude fewer parameters (5%∼10%). It also performs competitively against the large monolingual models.

The effects of transliteration: To understand the impact of orthography, we also fine-tuned the multilingual models on data transliterated from Ge'ez script to ASCII following a standard mapping (Gasser, 2011). In alignment with Muller et al. (2021), this resulted in a dramatic increase in performance for mBERT, but it had a slightly negative effect on XLM-R. We believe the reason lies in the tokenization schemes of the two models: mBERT relies on character-level word pieces, while XLM-R, following RoBERTa, uses byte-level byte-pair-encoding, making it possible for the former to come across unseen characters during inference. The evaluation results and the number of parameters for each model are presented in Table 1.

4 Conclusions and Future Work

In this work, we develop monolingual PLMs for Tigrinya using a corpus that we compiled from news sources and compare their performances with popular multilingual models on two downstream tasks. The evaluation results show that the monolingual models with much fewer parameters perform better than or competitively to their large multilingual counterparts. As future work, we plan to explore a multilingual model trained on languages closely related to Tigrinya, such as Amharic and Tigre. Finally, we hope that the new corpus and trained models will have a positive impact on Tigrinya NLP as similar resources did on other languages.

Acknowledgement

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Clark, K., Luong, M., Le, Q. V. & Manning, C. 2020. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR.*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. 2020. *Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Devlin, J., Chang, M., Lee, K. & Toutanova, K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Ebrahimi, A. & Kann, K. 2021. *How to Adapt Your Pretrained Multilingual Model to 1600 Languages. ACL/IJCNLP.*

Gaim, F., Yang, W. & Park, J. 2021. *TLMD: Tigrinya Language Modeling Dataset. Zenodo, 2021, 7, https://doi.org/10.5281/zenodo.5139094.*

Gasser, M. 2011. *HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya.*

Jonathan Heek and Anselm Levskaya and Avital Oliver and Marvin Ritter and Bertrand Rondepierre and Andreas Steiner and Marc van Zee 2020. *Flax: A neural network library and ecosystem for JAX. http://github.com/google/flax.*

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. 2020. *The State and Fate of Linguistic Diversity and Inclusion in the NLP World. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis M., Zettlemoyer, L., Stoyanov, V. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.*

Muller, B., Anastasopoulos, A., Sagot B., & Seddah, D. 2021. *When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Rust, P., Pfeiffer, J., Vuli'c, I., Ruder, S. & Gurevych, I. 2021. *How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.*

Tedla, Y., Yamamoto, K. & Marasinghe, A. 2016. *Tigrinya Part-of-Speech Tagging with Morphological Patterns and the New Nagaoka Tigrinya Corpus. International Journal Of Computer Applications* 146 pp. 33-41 (2016).

Tela, A., Woubie, A. & Hautamäki, V. 2020. *Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya. ArXiv, abs/2006.07698.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. 2017. *Attention is all you need. Advances In Neural Information Processing Systems, pp. 5998-6008.*

Wang, Z., Karthikeyan, K., Mayhew, S. & Roth, D. 2020. *Extending Multilingual BERT to Low-Resource Languages. EMNLP FINDINGS.*

Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and R'emi Louf and Morgan Funtowicz and Jamie Brew 2020. *Transformers: State-of-the-Art Natural Language Processing. EMNLP.*