Genealogic Approach to Low-Resource Language Modelling

Brandon Chaperon
Student
McGill University
Computer Science and Linguistics
brandon.chaperon@mcgill.ca

Suheng Qian
Student
McGill University
Computer Science and Linguistics
suheng.qian@mail.mcgill.ca

Abstract

An open problem in computational linguistics is how to model concepts whose data is scarce. One potential area where this would be useful is in modelling minority languages, or even those spoken by a large population but that do not have much data. A variant of BERT for multiple languages is mBERT (Multilingual BERT), which can currently model 104 of the over 7000 languages in the world (Devlin et al., 2018; Devlin, 2019). We will be continuing upon the works of Wu and Dredze (2020), who compared the differences of performance for BERT pre-trained on languages that have an abundance of resources versus those with low amounts of resources. Multilingual BERT has a much better performance when it comes to high-resources languages, while for languages with relatively low resources, the performance is much worse. We plan on investigating whether using data from closely related languages can help bridge the gap between low and high resource languages. We will be focusing on *Wolof*, an Atlantic Niger-Congo language which mBERT is not already pre-trained on. This language has 5 million speakers but is ranked 256th in number of Wikipedia articles. We compare the performance change after pre-training mBERT on *Yoruba* and *Swahili*, both in the Niger-Congo family, and *Amharic*, a Afro-Asiatic (Semitic) language. We conclude that closely related languages (i.e. Yoruba and Swahili) can help improve performance on tasks that are dependent on syntactic information. This technique can be adopted when limited amounts of data are available for that language. On the other hand, using an unrelated language can potentially impede the model.

1 Introduction

There has been much research done on using machines to read, understand, and generate language. With the age of big data, it has become increasingly easier to have access to large samples of text to use to train models. However, there is an unbalanced amount of data accessible for each language. Most research is done on languages like *English* and *Mandarin Chinese*, who's features unfortunately do not account for even a small fraction of those present in languages all over the world. Some languages, referred to as "low-resource", do not have the same opportunity of being well documented. This can be due to a low number of speakers, their culture, or for political and economic reasons. With more than 7000 languages in the world, a meagre 104 in mBERT barely touches upon 1.5% of all languages. Due to this reason, models perform much worse on modelling low-resource languages. This remains an open problem in computational linguistics, which we intend on further investigating.

We will attempt to find some potential solutions to the underperformance of BERT-based models on low-resource languages and provide NLP tools which will help further research in this area. Some potential challenges include: finding enough data to train our models (there are some languages that do not even have their own Wikipedia pages) and alternative tokenization techniques for languages that do not use the same writing system that we're familiar with (e.g. logographic or mixed-writing systems). Therefore, we needed to find ways to use data efficiently and get a sense of how to use small datasets whilst modelling a language accurately. We will be focusing on testing the models' syntactic understanding, which we believe is the basis for getting closer to understanding and using language (Chi et al., 2020). A good task to evaluate this on is part-of-speech (POS) tagging, which relies heavily on grammatical understanding. We will compare this task to named-entity recognition (NER) since this task relies less on syntax. The syntax from other related languages should not help the lower resource model as much.

A study by Wu and Dredze (2020) showed the disparity of performance between high-resource languages and low-resource languages. Therefore,

we would like to conduct a similar study to improve and test the performance of mBERT. Similarly to Goldberg (2019), who used tests like subject-verb agreement, we will be testing the performance of the models on some other syntactic phenomena. As stated, we will be testing our models on part-of-speech (POS) tagging and named-entity recognition (NER). Not all languages function in the same ways, so it is challenging to find unbiased methods to measure and compare the performances on each of them.

In addition, similar to Linzen and Baroni (2021) and Goldberg (2019), who used tests like subject-verb agreement, we will be testing the performance of the models on some other syntactic phenomena. Not all languages function in the same ways, so it will be challenging to find unbiased methods to measure and compare the performances on each of them.

Thus, our paper will attempt to find better ways to (1) augment the performance of low-resource languages (LRLs) on linguistic tasks and (2) compare standard syntactic understanding across these languages. We will be testing on *Wolof*, an Atlantic Niger-Congo language on mBERT base, and on mBERT fine-tuned on *Yoruba*, *Swahili*, and *Amharic*. We believe these constitute a relatively different array of language family relations. We intend on further investigating how models like BERT can get us closer to connecting the world one language at a time.

2 Related Work

Wu and Dredze (2020) have found significant disparities between high-resource languages and low-resource languages on both monolingual BERT (base and large) and mBERT. Low-resource languages have even worse performance in monolingual BERT. They compared the performance on a Name Entity-Recognition (NER) task and concluded that, the larger the task-specific supervised dataset, the better the downstream performance on NER. Statistical analysis shows a correlation between the amount of resources and mBERT performance but can not give a causal answer on why high-resource languages within mBERT perform poorly too.

They also compared the performance between monolingual BERT and mBERT, by implementing an experiment using four low-resource languages. They found that monolingual BERT performs better than mBERT on all metrics with the exception of Latvian Part-of-Speech (POS). One of the insights they found by comparing four low resource languages on monolingual BERT and mBERT is that mBERT might be able to adapt features from the training data of one language to perform tasks on another. Similarly, Goldberg (2019) compared BERT-base against BERT-large and found that BERT-base actually performs better than BERT-large on many syntactic conditions. Both of these findings suggest that in order to improve the performance on low-resource languages, a multilingual model is an optimal option and we should focus on the quality of the dataset and the use of its data, not necessarily on the size of the model.

2.1 Augmenting data

In Yang et al. (2020), they proposed an alternative pre-training method to the already established masking one called Code-Switching Pre-training (CSP). Although applied for Neural Machine Translation (NMT), this is still a relevant method for us to test on low-resource language models. Instead of randomly masking words in sentences, they replaced these words with their translation obtained from a high-resource language. In this way, CSP is able to pre-train the NMT model by explicitly making the most of the cross-lingual alignment information extracted from the source and target monolingual corpora. This method produced a great improvement over other pre-training methods. Since it is known that mBERT has the ability to learn the structure of a language by using data from other languages (Wu and Dredze, 2020), this method can be adopted to improve the performance of a model on low-resource languages. They found that we can benefit from pairing linguistically related languages. So we will not only augment the data from the chosen languages, but also from related languages.

2.2 Comparing cross-linguistic syntactic understanding

Similarly to Goldberg (2019), we will use tests of the same kind as subject-verb agreement to evaluate the syntactic competence of these models. However, we must first find which tests work cross linguistically to be able to reliably compare their performances. In Chi et al. (2020), they showed that subspaces of mBERT representations recover syntactic tree distances in languages other than English and that these subspaces are approximately shared

across languages. The study provided evidence mBERT learns representations of syntactic dependency labels, in the form of clusters that largely agree with the Universal Dependencies taxonomy. This evidence suggests that even without explicit supervision, multilingual masked language models learn certain linguistic universals. This shows us it will be possible to find good cross linguistic syntactic tests.

2.3 Comparing BERTs

Similarly to Goldberg (2019), we will use tests of the same kind as subject-verb agreement to evaluate the syntactic competence of these models. However, our tests must work cross linguistically to be able to reliably compare their performances. In Chi et al. (2020), they showed that subspaces of mBERT representations recover syntactic tree distances in languages other than English and that these subspaces are approximately shared across languages. The study provided evidence mBERT learns representations of syntactic dependency labels, in the form of clusters that largely agree with the Universal Dependencies taxonomy. This evidence suggests that even without explicit supervision, multilingual masked language models learn certain linguistic universals. This shows us it will be possible to use good cross linguistic syntactic tests. Thus, we have chosen to test our models on part-of-speech (POS) tagging and compare it to named entity-recognition (NER).

3 Datasets

So far mBERT has been trained on 104 languages. The top 100 languages with the most articles on Wikipedia were selected, plus some extra ones. Google has pre-trained mBERT on these languages' Wikipedia pages.

To test the change in performance of our model on part-of-speech (POS) tagging, we used Universal Dependencies' annotated data. Their datasets contain annotated "parts of speech (POS), morphological features, and syntactic dependencies". We used the Wolof (Wolof-WTB) dataset to test the model fine-tuned on other languages. To get a baseline for a high-resource language, we used an English (UD-English-LinES) treebanks also obtained from the Universal Dependencies.

For named entity-recognition (NER) we used Masakhane's NER dataset (Adelani et al., 2021). We did the same as with POS tagging and only used Wolof's dataset to test on different fine-tuned models on different languages. Thankfully, due to Masakhane's great strides, we were able to find labelled data for African languages. However, since the dataset was specific to these languages, we could not compare to a baseline on English.

4 Evaluation Metrics

We used similar metrics from a paper by Wu and Dredze (2019). The two evaluation metrics were part-of-speech (POS) tagging and named entity-recognition (NER). Universal Dependencies have treebanks for over 100 languages containing POS tagging and Masakhane have NER labelled data for 10 African languages. Thus, for each task, we trained a task-specific model on a base mBERT along with mBERT fine-tuned individually on each language.

For both of these evaluation methods, we only used 500 sample sentences for both languages. We split our data in 75% training (375 samples) and the remaining ones for testing (125 samples).

5 Models

For the baseline, we used multilingual BERT-cased because the models we are comparing to are based on a cased mBERT. We trained our baselines separately with English and Wolof. Our NER dataset is extremely unique because it only contains African (and low-resource) languages. Therefore, we were not able to find an English dataset with the same format. In order to compare them with our baselines, we used multilingual BERTs (BERT-base-multilingual-cased) that were enhanced by Adelani et al. (2021). They are all fine-tuned with extra data from specific languages we used to test our hypothesis (Yoruba, Swahili, and Amharic). We fine-tuned them for POS tagging and NER recognition by adding a linear layer with an input dimension of 768. The output dimension was the size of the number of label tags. For POS tagging we had 18 tags and for NER we used 10 tags (including the <*pad*> tag). For loss, we used Cross Entropy and optimised using Adam with a learning rate of 1E-6.

6 Results

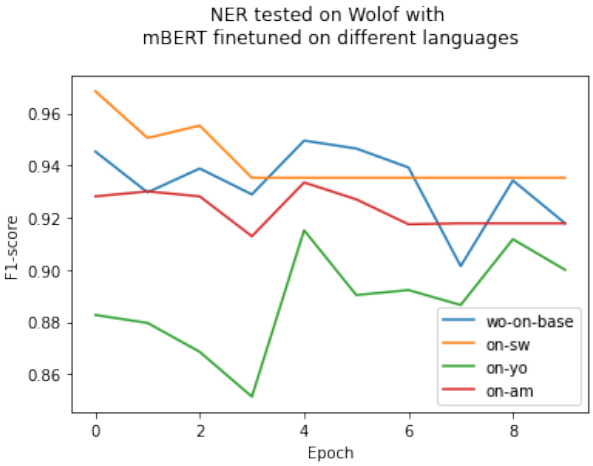
The original fine-tuned models by (Adelani et al., 2021) improved the NER f1 scores from 86.80 to 89.36 for *Swahili*, 78.97 to 82.58 for Yoruba, and also improved the Amharic performance (starting from 0 because data was non-existent beforehand).

Rank	Language	Local Name	Family	Speaker
256	Wolof	Wolof	Niger-Congo (Atlantic)	5 million
107	Yoruba	Yorùbá	Niger-Congo (Volta-Congo)	38 million
84	Swahili	Kiswahili	Niger–Congo (Volta-Congo > Bantoid)	90 million
136	Amharic	Amarñña	Afro-Asiatic (Semitic)	21 million

Table 1: Basic information for the languages involved in this study
[a] List of Wikipedias, [b] List of Wikipedias by speakers per article

on all languages did not impact the performance in any significant way.

Our experiments have shown that the performance on POS tasks for *Wolof* with the *Swahili* and *Yoruba* fine tuned models improved from 83.37 to 84.54 and 84.14 on average respectively. On the other hand, for Amharic, the f1 score improved to only 84.03, but did worse on average. These results demonstrated that closely related languages improve the performance on POS tasks but non-related languages do not help at all.



In addition, more experiments showed that the performance on NER tasks for *Wolof* on all languages did not impact the performance in any significant way.

7 Discussion

First, we must give a linguistic background of the language chosen to be used. *Wolof*, *Yoruba* and *Swahili* are all contained within the Niger–Congo language family. *Yoruba* and *Swahili* can be more specifically categorised into the Volta-Congo (and *Swahili* in Bantoid) family. Although spoken in the same continent, Wolof and Amharic are in different language families (Atlantic–Congo and Semitic respectively) so it is expected to get a worse performance using a model fine-tuned on it than with the previously mentioned ones.

In terms of POS-tagging, both Swahili and Yoruba produced visible improvements compared to the original model. This conclusion is drawn by comparing the improvements of a fine-tuned model on its own language and the improvements we discovered in our studies. This suggests that if we want to improve the performance of mBERT on a certain low-resource language that we currently do not have much data for, we can use related languages. POS-tagging relies heavily on syntactic properties of language. Combining data from other languages in the same or similar language families might help to improve the performance of the model. However there is a limitation to this: language isolates like Haida and Basque do not have related languages available to use this approach.

We hypothesised that Amharic POS-tagging would not produce a better result compared to the original model because Amharic does not share a proto-language with Wolof, the language we are testing on.

As for NER, (Poibeau, 2001) argue that NER fine tuning does not cross over to other domains well. We extend this notion to cross linguistic performance. NER performance should stay language specific and not cross over to other related languages. Since NER does not directly rely on syn-

Language	POS	NER
Wolof	83.37	93.31
Swahili	84.54	94.93
Amharic	84.03	92.30
Yoruba	84.14	88.78
English	89.62	N/A

Table 2: Average F1 score for POS-tagging and NER

tactic structures of the language, we did not expect to get much variation whether we used related languages or not.

Both of these conclusions suggest that our hypothesis only works when it comes to tasks that are highly syntactic-dependent. We argue that this is due to the classification of language families, which are mostly based on the similarities of syntactic structures. mBERT appears to contain a syntactic understanding of language and this innate knowledge can be used to improve cross-linguistic performance. Here we showed that this property can be used to augment modelling of low-resource languages.

8 Conclusion

Fine-tuning multilingual BERT (using masking) on a language will potentially improve its performance on another language within the same language family. More specifically, this applies to tasks that are syntactic-dependent (like part-of-speech tagging). For tasks that are not dependent on syntactic context (For example, named entity-recognition), it did not improve the outcome. In some instances, it could even worsen the results. This method provides a shortcut for researchers that do not have the resources or equipment available to pre-train an entire model from scratch. More importantly, low-resource languages that do not have much dataset available can still be modelled using closely related languages.

For future studies, there is a potential application for Transformer models in the study of historical linguistics. Linguists can use these methods to test their hypothesis about language genealogy classification. They can explore the syntactic relationship between different languages and potentially classify them within the same language family.

The code to our project can be found here

9 Contributions

Both members contributed equally to the whole of the project.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Jacob Devlin. 2019. Bert/multilingual.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Leila Kosseim Poibeau, Thierry. 2001. Proper name extraction from non-journalistic texts. In *Language and Computers*, Online.

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

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.