Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2475–2478
Marseille, 11–16 May 2020

Adaptation of Deep Bidirectional Transformers for Afrikaans Language

Sello Ralethe
Rand Merchant Bank
Johannesburg, South Africa
Ralethe.sello@gmail.com

Abstract

The recent success of pretrained language models in Natural Language Processing has sparked interest in training such models for languages other than English. Currently, training of these models can either be monolingual or multilingual based. In the case of multilingual models, such models are trained on concatenated data of multiple languages. We introduce AfriBERT, a language model for the Afrikaans language based on Bidirectional Encoder Representation from Transformers (BERT). We compare the performance of AfriBERT against multilingual BERT in multiple downstream tasks, namely part-of-speech tagging, named-entity recognition, and dependency parsing. Our results show that AfriBERT improves the current state-of-the-art in most of the tasks we considered, and that transfer learning from multilingual to monolingual model can have a significant performance improvement on downstream tasks. We release the pretrained model for AfriBERT.

Keywords: Afrikaans, BERT, AfriBERT, Multilingual

1. Introduction

Pretrained language models use large-scale unlabelled data to learn highly effective general language representations. These models have become ubiquitous in Natural Language Processing (NLP), pushing the state-of-the-art in a large variety of tasks involving natural language understanding and generation such as natural language inference, sentiment classification, and semantic textual similarity. A few prominent examples of these models include GPT (Radford et al., 2018), ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNET (Yang et al., 2019). An advantage that these pretrained language models have over traditional task-specific approaches is that they can be trained in an unsupervised manner. However, they can be difficult to implement because of the amount of data and computational resources needed for pretraining.

The need for large-scale data has limited the availability of these pretrained models mostly to the English language. For other languages, these models are implemented as multilingual models. Lample and Conneau (2019) have demonstrated that multilingual models perform poorer than their monolingual models.

In this paper, we show how a monolingual model can be trained using multilingual initialisation. We use multilingual BERT, with Afrikaans as a target language for transfer.

To train a monolingual language model for Afrikaans, we used the following corpora:

- OSCAR (Ortiz Su´arez et al., 2019)
- Afrikaans News Article
- NCHLT Afrikaans Text Corpora (Puttkammer et al., 2014)
- Afrikaans Novels (Project Gutenberg)
- Afrikaans Wikipedia
- OPUS (Tiedemann, 2012)

We train the model following the BERT architecture. We evaluate our trained model, called AfriBERT, on the following downstream tasks for Afrikaans: part-of-speech (POS) tagging, named-entity recognition (NER), and dependency parsing. AfriBERT improves the state-of-the-art for most of these tasks when compared with multilingual approach.

In summary, we make three main contributions:

- We use transfer learning to train a monolingual BERT model on the Afrikaans language using a combined corpus.
- We evaluate AfriBERT on three downstream tasks, and improve the state-of-the-art results in all tasks, confirming the significant improvement of transfer learning from multilingual to monolingual.
- We open-source AfriBERT as part of the HuggingFace's Transformers Library (Wolf et al., 2019), and also on our GitHub repository: https://github.com/sello-ralethe/AfriBERT

2. Background work

Bidirectional Encoder Representation from Transformers (BERT) is a deep contextual representation based on a series of transformers that are trained by a self-supervised objective (Devlin et al., 2019). Unlike other language models such as GPT and ELMO, BERT is trained by the Cloze task (Taylor, 1953), also referred to as masked language modelling, instead of right-to-left or left-to-right language modelling. This allows BERT to freely encode information from both directions in each layer. In addition, BERT also optimizes a next sentence classification objective (Devlin et al., 2019).

In the pre-training phase, BERT makes use of two objectives: masked language model (LM) and next sentence prediction (NSP). In masked LM, some input tokens are randomly masked, and BERT then tries to predict these masked tokens (Devlin et al., 2019).

For the NSP task, BERT aims to predict the next sentence given a certain sentence. At training time, 50% of the paired sentences are consecutive sentences while the rest of the sentences are paired randomly. The purpose of adding the NSP objective is that many downstream tasks such as question answering and language inference require sentence-level understanding, which is not directly captured by LM objectives (Devlin et al., 2019).

Since the publication of Multilingual BERT (henceforth, mBERT) (Devlin et al., 2019), there has been numerous publications of positive experimental results for various

multilingual tasks, such as natural language inference and part-of-speech tagging (Pires et al., 2019).
mBERT follows the same model architecture and training procedure as BERT. mBERT is trained on Wikipedia data of 104 languages with a shared word piece vocabulary that contains 110k subwords calculated from the WordPiece model (Wu et al., 2016). Exponentially smoothed weighting was applied to prevent high-resource languages from dominating the training data.
During training, there's no use of any explicit cross-lingual signal, or mechanism to encourage translation equivalent pairs to have similar representations. Also, there is no use of markers to denote the input language.
In mBERT, the WordPiece modelling strategy allows the model to share embeddings across languages. To account for varying sizes of Wikipedia training data in different languages, training uses a heuristic to subsample or oversample words when running WordPiece as well as sampling a training batch, random words for cloze and random sentences for next sentence classification.

3. Model Architecture

Our approach is based on BERT (Devlin et al., 2019), which is a multilayer bidirectional Transformer (Vaswani et al., 2017). For more depth details about Transformers, we refer the reader to Vaswani et al. (2017).
For AfriBERT, we used the original $BERT_{BASE}$ configuration: 12 layers (Transformer blocks), 12 attention heads, and 768 hidden dimensions. This amount to 110M parameters.
We trained task-specific layers in accordance with Devlin et al. (2019).

4. Language Transfer

In this paper, we consider how we can use transfer learning to go from multilingual BERT to monolingual AfriBERT. According to Devlin et al. (2019), BERT uses the subword segmentation algorithm to handle the problem of vocabulary size. In a multilingual language model, only a small part of the entire vocabulary is used for a single language. This implies that tokenization in a multilingual model produces longer sequences than in a monolingual model. The design of the Transformer model compels us to steer away from longer sequences to minimise computational complexity.
For our study, we investigated the feasibility of initialising a monolingual model using a multilingual model. To implement this, we use the acquired knowledge of the target language which is already captured during training of the multilingual model. Literature shows that we can improve performance of a model when we use data from multiple languages to train that model (Mulcaire et al., 2018). We make use of mBERT to initialise all parameters of our model, except word embeddings.
We used subword-nmt to construct a subword vocabulary for our model. To train the subword vocabulary, we used the corpora that comprises of texts from multiple sources. This resulted in a new monolingual Afrikaans subword vocabulary, containing longer words and subwords than in a multilingual one.
Our word embedding matrix is a result of gathering monolingual embeddings from those of mBERT. In the resulting matrix, we did not change embeddings of all tokens present in both mBERT and our model. We did the same for special tokens such as [UNK] and [CLS]. Tokens that were present in mBERT and not in our model were replaced by those in our model vocabulary. These tokens are mostly longer subword units obtained by combining shorter units present in both mBERT and our model.
The model with reassembled vocabulary and embeddings matrix was trained on the same data that was used for constructing the monolingual vocabulary. Following (Liu et al., 2019), we use Adam (Kingma and Ba, 2014) to optimise the model, ($\beta_1 = 0.9,\ \beta_2 = 0.98$) for 100k steps. We use batch sizes of 200 sequences. Each sequence contains complete sentences.

5. Evaluation

5.1 Part-of-speech tagging and dependency parsing

We start by evaluating AfriBERT on part-of-speech (POS) tagging and dependency parsing. The part-of-speech of a word is determined by the morpho-syntactic behaviour of the word in the specific context. This involves assigning corresponding grammatical category to each word. Dependency parsing involves the prediction of labelled syntactic tree that captures the syntactic relations between words. Unlike POS tags, dependency relation tags are hierarchical. This makes it possible for us to specify dependencies between words even when the nature of that dependency is not captured accurately in one of the more specific categories.
We run our experiments using the NCHLT Afrikaans Text Corpora (Puttkammer et al., 2014) and Afribooms Afrikaans Dependency Treebank (Augustinus et al., 2016). NCHLT Afrikaans Text is a collection of text documents from the South African government domain crawled from gov.za websites and collected from various language units. The Afribooms Afrikaans Dependency Treebank (Augustinus et al., 2016) includes annotations for lemma, POS and dependency relations. This corpus contains manually annotated POS tags, and dependency relations.
We evaluate the performance of AfriBERT using the standard UPOS accuracy for POS tagging, and Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS) for dependency parsing.

5.2 Named Entity Recognition

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organisations, locations, etc. Identifying these categories is an essential tool in the development of complex information extraction and retrieval tools. We use the NCHLT Text Resource Development for named entity. The named entity annotated data is a combination of the 50k tokens annotated during the NCHLT text resource development project, as well as additional data from the NCHLT text corpora previously collected.

The annotated set consists of a minimum of 15k tokens annotated as one of the following phrase types:

- ORG – Organisation
- LOC – Location
- PER – Person
- MISC – Miscellaneous
- OUT – not considered part of any named entity

For NER, we report the 3 metrics that are commonly used to evaluate models: precision, recall, and F1 score. Precision measures the percentage of entities found by the system that are correctly tagged, recall measures the percentage of named entities present in the corpus that are found, and F1 score combines both precision and recall measures to give a general idea of how the model performs.

6. Experiments

In this section, we measure the performance of AfriBERT by evaluating it on the three tasks: POS tagging, dependency parsing, and NER.

6.1 Experimental Setup

For each task, we append the relevant predictive layer on top of AfriBERT's Transformer architecture.
We use the same sequence tagging architecture as Devlin et al. (2019). We tokenize the input sentence, feed it to AfriBERT, get the last layer's activations, and pass them through a final layer to make predictions.
We fine-tune AfriBERT independently for each task and each dataset.

6.2 Results

For POS tagging and dependency parsing, we compare AfriBERT to mBERT in Table 1. All reported results are obtained by averaging across 5 runs.

Model	NCHLT Afrikaans Text Corpora			Afribooms Afrikaans Dependency Treebank		
	UPOS	UAS	LAS	UPOS	UAS	LAS
mBERT	97.34	91.62	89.72	98.71	93.33	91.81
AfriBERT	98.56	92.43	91.28	99.12	95.65	94.47

Table 1: POS and dependency parsing scores of mBERT and AfriBERT. Here mBERT was fine-tuned in the same conditions as AfriBERT.

The results in Table 1 show that AfriBERT has a higher performance than mBERT in POS tagging and dependency parsing tasks. We observe a better performance of AfriBERT on the Afribooms Afrikaans Dependency Treebank in both parsing and tagging tasks.
For named entity recognition, our results in Table 2 show that AfriBERT achieves a significantly better precision than mBERT. We observe the same for Recall and F1 score.

Model	F1	Precision	Recall
mBERT	81.41	80.19	83.25
AfriBERT	85.46	87.64	85.84

Table 2: Results for NER on the NCHLT Afrikaans Text Corpora

6.3 Discussion

The results of our experiments show that AfriBERT has a better performance compared to mBERT for the three downstream tasks we considered. Our results confirm the hypothesis that pretrained language models can be effectively fine-tuned for various downstream tasks, as observed for the English language in previous work. Our work also shows that monolingual models significantly outperform their multilingual counterparts. We note here that the size of data used for AfriBERT was possibly crucial to the observed performance. In addition to Wikipedia data used for Afrikaans in mBERT, our data had addition text from multiple sources. The availability of additional data provides more diversity in the pretraining distribution.

7. Conclusion

We presented AfriBERT, a language model for the Afrikaans language. We have shown here that AfriBERT significantly improves performance on several Afrikaans NLP tasks compared to mBERT. Also, our results show that initialising a monolingual model from a multilingual one results in even better improvements. Lastly, AfriBERT is lighter than mBERT and other BERT-based approaches such as XLM. We hope that opensourcing AfriBERT will serve as baseline for future research in NLP for Afrikaans.

8. Acknowledgements

We thank Tensorflow Research Cloud program for providing us with computing resources for training our model.

9. Bibliographical References

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACLHLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Kingma, D. P. and Ba, J., L. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mulcaire, P., Swayamdipta, S., and Smith, N. (2018). Polyglot semantic role labelling. *arXiv preprint arXiv:1805.11598.*

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep

contextualized word representations. In Walker et al. (Walker et al., 2018), pages 2227–2237.
Pires, T., Schlinger, E., Garette, D. (2019). How multilingual is multilingual bert? In Proceedings of Association of Computational Linguistics (ACL).
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Languagemodels are unsupervised multitask learners. *OpenAI Blog*, 1:8.
Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quaterly, 30, 415-433.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Isabelle Guyon, et al., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J (2019). HuggingFace's Transformers:State-of-the-art Natural Language Processing. *arXiv preprint arXiv:abs/1910.03771*
Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR*, abs/1904.09077.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

10. Language Resource references

Augustinus, L., Dirix, P., van Niekerk, D., Schuurman, I., Vandeghinste, V., Eynde, F. and Van Huyssteen, G. (2016). AfriBooms: An Online Treebank for Afrikaans.
Project Gutenberg. (n.d.). Retrieved October, 2016, from www.gutenberg.org.
Puttkammer, M., Schlemmer, M. and Bekker, R. (2014). Afrikaans NCHLT Annotated Text Corpora. South African Language Resource Management Agency, Potchefstroom, 1.0, ISLRN 139-586-400-050-9.
Ortiz Su´arez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.
Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*