EFFECT OF TOKENISATION STRATEGIES FOR LOW-RESOURCED SOUTHERN AFRICAN LANGUAGES

Jenalea Rajab
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
jenalea.rajab@gmail.com

ABSTRACT

Research into machine translation for African languages is very limited and low-resourced in terms of datasets and model evaluations. This work aims to add to the field of neural machine translation research, for four low-resourced Southern African languages. The effect of two byte pair encoding tokenisation algorithms (subword_nmt and SentencePiece), with various parameters, are evaluated. The paper builds upon previous research in the field for comparison, using an optimised transformer architecture and pre-cleaned data to translate English to Northern Sotho, Setswana, Xitsonga and isiZulu. The results obtained show improvements in the previous BLEU scores obtained for Setswana and isiZulu.

1 INTRODUCTION

There has been a slight increase in the focus of machine translation for low-resourced languages in recent years, most notably in African languages through the collaboration of African curators, annotators and language technologists (Abbott, 2021). However there is still a large amount of work to be done in the field, with varying approaches and optimal models to be tested in the setting of low and unconsolidated resources.

This research evaluates two different Byte Pair Encoding (BPE) algorithms (subword_nmt and SentencePiece) using an optimised transformer architecture. The models are trained on four agglutinative (Zerbian, 2007) Southern African Bantu languages (Mesthrie, 2002), namely Northern Sotho, Setswana, Xitsonga and isiZulu. The work builds directly on research, into Neural Machine Translation (NMT) for low-resourced African Languages, by Martinus & Abbott (2019) and Biljon et al. (2020). The results show increased performance for Setswana and isiZulu when using SentencePiece BPE tokenisation, in comparison to the previous work on the same dataset.

Section 2 describes previous related work and the inspiration for this research. The methodology including the datasets, tokenisation strategies and trained models are described in Section 3. The results are presented and analysed in Section 4 and 5 respectively. A impact statement for this work is outlined in Section 6, followed by the Conclusion.

2 BACKGROUND

In 2019, Martinus & Abbott (2019) wrote a position piece on the problems facing NMT for low-resourced African languages. They highlighted a number of problem areas, including lack of research into adequate models. To address this, their research provided results from the implementation of NMT models, using the transformer architecture, for five Southern African languages from the Autshumato dataset (Groenewald & Fourie, 2009). Additionally, they provided an ablation study on the BPE token size, for each of the languages, using the subword_nmt algorithm. In the ablation study they varied the token size (5k - 40k) implemented, showing that it directly affected the model's final performance, and determined an optimal token size for each language. These results are discussed in Section 3.2.1 .

Following this, Biljon et al. (2020) expanded on the above research by testing the optimal transformer depth for use on low-resourced languages. Making use of the same dataset, but only analysing a subset of the five languages, they determined that medium-depth transformers worked optimally for the low-resource language translation models; and provided their open source code [1] along with the implementation results.

Both of these studies made use of BPE tokenisation, specifically using the subword_nmt BPE algorithm. BPE has shown good results when used for low-resourced languages as it helps with rare and out-of-vocabulary words, which is common in the low-resource setting (Sennrich et al., 2015). The objective of this study is to further expand on the work of Martinus & Abbott (2019) and Biljon et al. (2020) in determining if the SentencePiece BPE tokenisation strategy will assist performance for these low-resourced languages. SentencePiece BPE encodes white space and therefore does not require the words in the corpus to be white-space separated. Since the Bantu languages tested are agglutinative, by not requiring the text to be 'pre-tokenised' it could improve translation performance (Zerbian, 2007) (Horan).

Currently, to the best of the author's knowledge, an in-depth comparison has not be done between subword_nmt BPE and SentencePiece BPE tokenisation on African languages, therefore this research aims to support further expansion of African Natural Language Processing.

3 METHODOLOGY

In order for comparability to the previous work described, the pre-processed Autshumato (Groenewald & Fourie, 2009) dataset from Martinus & Abbott (2019) and the open-source optimised transformer implementation by Biljon et al. (2020) is used to train the NMT models for each language. The data processing, data splits, transformer architecture and hyper-parameter in settings in Biljon et al. (2020) are kept the same for direct result comparisons. Only the tokenisation strategy has been modified to include the SentencePiece BPE algorithm and is compared to the existing subword_nmt BPE implementation. The models are trained to perform English to Target Language translation for the following Bantu languages:

- Northern Sotho
- Setswana
- Xitsonga
- isiZulu

The SentencePiece BPE implementation is tested using various vocabulary sizes (4k, 8k and 16k), while the subword_nmt algorithm is evaluated using only the optimal token size for each language respectively (determined from the ablation study by Martinus & Abbott (2019)). An overview of the described methodology is shown in Figure 1. The code for the below experiments have been published on GitHub (`https://github.com/JenaleaR/Tokenisation_African_Languages`) and have been made open-source to promote reproducibility, and enable further research.

3.1 LANGUAGES

The target languages evaluated are agglutinative Southern Bantu African languages (Mesthrie, 2002)(Zerbian, 2007). All the languages are similar in structure and vocabulary (Herbert & Bailey, 2002), with a subject-verb-object word order and a rich noun class system Zerbian (2007). Additionally the Bantu languages share root words (Mills, 2005), with Setswana and Northern Sotho being mutually-intelligible (Martinus & Abbott, 2019).

3.2 DATA

The Autshumato dataset is a public parallel corpora of South African government data created by Groenewald & Fourie (2009) for NMT systems. The dataset contains aligned sentences for the four languages trained on (Martinus & Abbott, 2019).

[1] `https://github.com/ElanVB/optimal_transformer_depth`

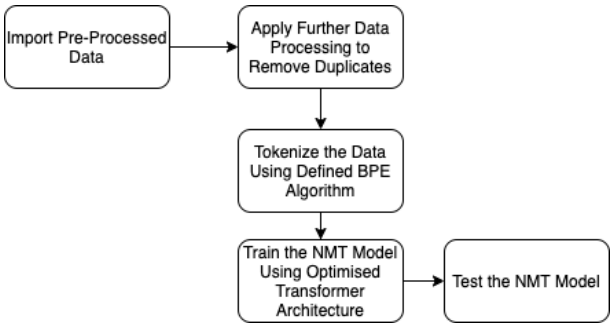

Figure 1: Methodology Overview

The dataset was pre-processed by Martinus & Abbott (2019), where duplicate sentences were found and removed along with their translations. Biljon et al. (2020) further processed the data, removing translation duplicates as well as 'almost duplicates', and ensured test data was filtered from the training and dev sets to avoid data leakage. 3000 parallel sentences were used for testing, the data was shuffled to remove bias and 1000 parallel sentences were set aside for validation, with the rest being used for training the models. The summary statistics of the post-processed sentences per language and their respective splits are shown in Table 1.

Table 1: Summary statistics of post-processed sentences per language

Target Language	**Training Sentences**	**Dev Sentences**	**Test Sentences**
Northern Sotho (nso)	20 672	1000	3000
Setswana (tn)	75 070	1000	3000
Xitsonga (ts)	134 823	1000	3000
isiZulu (zu)	17 292	1000	3000

3.2.1 SUBWORD BPE

Subword_nmt is a BPE tokenisation method that is defined by the number of merges (Sennrich et al., 2015). The algorithm separates the corpus by white space 'into words', counts the all pairs of adjacent characters, merges the characters of the most frequent pair and adds it to the vocabulary (Jurafsky & Martin, 2020). This process is repeated for the defined number of merges/tokens (Jurafsky & Martin, 2020).

In the work by Martinus & Abbott (2019) they noted initial experimentation showed that the merge/token number affected the model's final performance. To test this they performed an ablation study using subword_nmt and showed the optimal number of tokens for each language. Their results are listed in Table 2.

Table 2: Optimal number of subword_nmt BPE tokens for each language

Target Language	**BPE tokens**
Northern Sotho (nso)	4000
Setswana (tn)	40 000
Xitsonga (ts)	20 000
isiZulu (zu)	4000

Based on these results, the models in this research were trained using subword_nmt tokenisation, with the optimal token number per language, to provide a baseline evaluation before training the models using SentancePiece BPE tokenisation.

3.2.2 SENTENCEPIECE BPE

Subword_nmt requires the input text to be word tokenised in order to perform the merge operations, however since not all languages are space segmented between words this process could hinder the training and performance of NMT models, specifically on agglutinative languages (Horan). The SentencePiece BPE method by Kudo & Richardson (2018), addresses this problem by treating the input text as a Unicode character sequence, including the white spaces (Horan). White spaces are encoded with the meta symbol '_', as part of a token, which assists in decoding and makes the algorithm language agnostic (Horan). SentencePiece BPE uses the final vocabulary size as its parameter, differing from subword_nmt which requires the defined number of merge operations (Kudo & Richardson, 2018), this is done since SentencePiece also supports other segmentation algorithms (e.g. Unigram, Char and Word) (Kudo & Richardson, 2018).

The models are trained using the SentencePiece BPE tokenisation algorithm with varying vocabulary sizes (4k,8k and 16k), to determine if encoding the white space could improve performance for the agglutinative languages tested.

3.3 ALGORITHM

The medium-depth transformer architecture from Biljon et al. (2020) was used to train the translation models for each language, using the different tokenisation strategies. The tuned hyper-parameters where kept exactly as they had defined them. The medium transformer consists of 6 transformer layers (3 encoder and 3 decoder) with the learning rate set to 0.0003 and a batch token size of 4096. Beam search with a width of 5 was used to decode the test data. The transformer was implemented using the JoeyNMT toolkit, which is based on PyTorch and provides NMT features with simple adaptable code (Kreutzer et al., 2019). JoeyNMT supports BPE subword encoding learned with subword_nmt as well as SentencePiece (Kreutzer et al., 2019).

A model for each language was first trained and tested using the subword_nmt algorithm, with the respective optimal token size, in order to get a baseline comparison. The SentencePiece BPE algorithm was then implemented and language models were trained for each vocabulary size (4k, 8k and 16k). Each model was trained for 30 epochs in order to determine the effect of each configuration, while limiting the computation time. Following the initial results, in order to further test the performance seen by using SentencePiece BPE tokenisation, the models were trained for 100 epochs using the optimal vocabulary size noted for each language.

4 RESULTS

The initial BLEU results after 30 epochs are shown in Table 3. It is seen that the SentencePiece BPE (sp_bpe) achieved very good results across all languages, surpassing the subword_nmt (sw_bpe) performance for the initial tests. The optimal results for each language are achieved for different vocabulary sizes, which aligns with Martinus & Abbott (2019) findings that the token number affects the model performance. The models for Northern Sotho, Setswana and isiZulu achieved higher BLEU scores for lower vocabulary sizes, while the model for Xitsonga (with the largest dataset) shows higher BLEU scores as the vocabulary size increases.

Table 3: BLEU scores calculated for each tokenisation strategy (30 epochs), for English to Target language translations on test sets

Target Language	**sw_bpe**	**sp_bpe4k**	**sp_bpe8k**	**sp_bpe16k**
Northern Sotho (nso)	14.53 (4k)	18.31	**18.42**	17.42
Setswana (tn)	26.79 (40k)	**29.89**	27.63	29.29
Xitsonga (ts)	34.58 (20k)	33.99	34.85	**35.82**
isiZulu (zu)	1.71 (4k)	**7.59**	5.00	4.65

Following the initial tests, the models were retrained for 100 epochs using the optimal vocabulary size determined. The BLEU results after 100 epochs are shown in Table 4 with a comparison to the

results achieved by Biljon et al. (2020) and Martinus & Abbott (2019). It is seen that the Northern Sotho and Setswana BLEU results surpass those achieved by Biljon et al. (2020), with the result for Setswana surpassing the performance seen in Martinus & Abbott (2019) as well. A four fold increase in the baseline isiZulu performance by Martinus & Abbott (2019) is also achieved, using the SentencePiece BPE method with a vocabulary size of 4000. The Xitsonga model and Northern Sotho models were unable to achieve the performance seen in Martinus & Abbott (2019), with the implemented vocabulary sizes. A higher BLEU result was expected for Northern Sotho since is closely related to Setswana (which achieved good results) as they are linguistically similar (Martinus & Abbott, 2019) and further investigation into the optimal vocabulary size needs to be done. The results confirm those in Martinus & Abbott (2019), where the model performance is also related to the number of parallel sentences in the corpus.

Table 4: Test BLEU scores calculated for optimal SentencePiece BPE vocabulary sizes, compared to previous literature

Target Language	**sp_bpe result**	Biljon et al. (2020)	Martinus & Abbott (2019)
Northern Sotho (nso)	23.54 (8k)	17.67	**24.16**
Setswana (tn)	**32.40** (4k)	30.49	28.07
Xitsonga (ts)	39.56 (16k)		**49.74**
isiZulu (zu)	**13.06** (4k)		3.33

5 ANALYSIS

The SentencePiece BPE tokenisation resulted in an overall increase in performance when compared with the same models trained by Biljon et al. (2020) using subword_nmt. Example model outputs for Setswana and isiZulu have been included in Appendix A for a qualitative comparison. An in-depth qualitative analysis of the model results still needs to be done to determine if the high BLEU scores achieved correspond to accurate translations. It is highlighted by Martinus & Abbott (2019) that there exists quality issues within the isiZulu data, including mismatched translations and spacing between letters, therefore despite achieving a big performance increase in the BLEU score the translations may be inaccurate. Additionally the BLEU metric might not be expressive enough to evaluate the performance of the agglutinative language translations and a deeper analysis will assist in confirming the results (Biljon et al., 2020).

The initial tests show convincing results, therefore further research includes implementing an ablation study on the vocabulary size for the different language models. Only a small subset of vocabulary sizes were tested, therefore more research will assist in determining/confirming the optimal size for each language. An ablation study will also help determine the relationship between the size of the vocabulary, size of the dataset and the linguistics of the language.

6 IMPACT STATEMENT

This work could assist in the accurate creation of NMT models for English to Southern African languages. An increase in accurate translations could assist in bringing Africa into the scientific conversion and reduce the access to information gap, particularly in the online setting (Martinus & Abbott, 2019). Advances in low-resource language translation has great benefits for education, providing information to people in their native language. Additionally it could assist in the accurate detection of hate speech and violence incitement, which often goes unnoticed in current large-scale NMT systems which do not perform adequately on low-resourced languages (Bender et al., 2021)

It is noted that the dataset used is taken only from South African government data and if used in production could misrepresent current social movements and will align to the existing political regime at the time the data was collected (Bender et al., 2021). Additionally the data has not been adequately curated to determine the biases that exist. If any bias exists it will be encoded into the model, which could result in harm towards specific groups, if used in production for text generation or classification (Bender et al., 2021).

7 CONCLUSION

The effect of two byte pair encoding tokenisation algorithms, subword_nmt and SentencePiece, with various parameters were compared for use in NMT models. The research builds upon previous research in the field, using an optimised transformer architecture and pre-cleaned data to translate English to Northern Sotho, Setswana, Xitsonga and isiZulu. The results obtained show improvements in the previous BLEU scores obtained for Setswana and isiZulu when using the SentencePiece tokenisation algorithm. It is noted that a qualitative analysis of the model results still needs to be done, to determine if the high BLEU scores achieved correspond to accurate translations. Additionally further work includes implementing an ablation study on the SentencePiece vocabulary size, to define the relationship between the size of the vocabulary, size of the dataset and the linguistics of the language. This work provides a basis for further investigation into using SentencePiece tokenisation for NMT, with optimised vocabulary size, for agglutinative African Languages to achieve better performance.

REFERENCES

Jade Z. Abbott. Africannlp. Presentation, 2021.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. On optimal transformer depth for low-resource language translation. *CoRR*, abs/2004.04418, 2020. URL https://arxiv.org/abs/2004.04418.

Hendrik J. Groenewald and Wildrich Fourie. Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain, May 14–15 2009. European Association for Machine Translation. URL https://aclanthology.org/2009.eamt-1.26.

Robert K. Herbert and Richard Bailey. *The Bantu languages: sociohistorical perspectives*, pp. 50–78. Cambridge University Press, 2002. doi: 10.1017/CBO9780511486692.004.

Cathal Horan. Tokenizers: How machines read. URL https://blog.floydhub.com/tokenization-nlp/#sentencepiece.

Daniel Jurafsky and James H Martin. Speech and language processing. Third Edition draft, December 2020.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey nmt: A minimalist nmt toolkit for novices. In *ACL Anthology*, pp. 109–114, 01 2019. doi: 10.18653/v1/D19-3019.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018. URL http://arxiv.org/abs/1808.06226.

Laura Martinus and Jade Z. Abbott. A focus on neural machine translation for african languages. *CoRR*, abs/1906.05685, 2019. URL http://arxiv.org/abs/1906.05685.

Rajend Mesthrie (ed.). *Language in South Africa*. Cambridge University Press, 2002. doi: 10.1017/CBO9780511486692.

Wallace Mills. History 316.1 africa in the nineteenth century, 2005. URL http://smu-facweb.smu.ca/~wmills/course316/his316.html.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

Sabine Zerbian. A first approach to information structuring in xitsonga/ xichangana*. pp. 1–22, 2007.

A APPENDIX: TRANSLATION EXAMPLES

Table 5: Example Model Translations for English to Setswana

Description	Text
Example 1:	
Source (en)	The selection of the cultivar is of utmost importance.
Reference (tn)	Tlhopo ya mofuta o o jwalwang ke kgang e e leng botlhokwa thata.
Hypothesis (tn)	Go tlhopha mofuta o botlhokwa thata.
Example 2:	
Source (en)	For example , researchers should earn much more for an article published in an international journal than for those published nationally.
Reference (tn)	Ka sekai , babatlisisi ba tshwanetse go amogela go le gontsinyana fa ba gatisitse athikele mo lekwalopakeng la boditšhabatšhaba go feta fa ba e gatisitse mo go la naga ya rona.
Hypothesis (tn)	Sekao , babatlisisi ba tshwanetse go amogela athikele e e phasaladitsweng mo lekwalopakeng la boditšhabatšhaba go gaisa bao ba phasaladitsweng mo nageng yotlhe .
Example 3:	
Source (en)	Domestic visitor
Reference (tn)	Moengselegae
Hypothesis (tn)	Baeng ba selegae

Table 6: Example Model Translations for English to Zulu

Description	Text
Example 1:	
Source (en)	They obligingly sang Shosholoza when we requested the customary Venetian accompaniment to our cruise.
Reference (zu)	Bakuthokozela ukucula uShosholoza ngesikhathi sicela ukuba kudlalwe umculo ohambisana nesiko laseVenisi ohambeni lwethu.
Hypothesis (zu)	Bathola iShosholoza ngesikhathi sicela ukuthi amasiko aseVenetian aphelekezela ukuhamba kwethu
Example 2:	
Source (en)	There were actually two Smits, one was the driver and the other one was in the crate with me.
Reference (zu)	Empeleni babebabili oSmits, omunye wayengumshayeli kanti lona omunye ubesebenza nami emakredini.
Hypothesis (zu)	Kwakunezinkomba ezimbili zoSmit, omunye wabashayeli kanye namanye alowo oyedwa.
Example 3:	
Source (en)	the foreign vessel will fish under South African regulations and permit conditions .
Reference (zu)	lomkhumbi wangaphandle uzothobela yonke imithetho yokudoba nesimiso sephomedi sase South Africa.
Hypothesis (zu)	umkhumbi wangaphandle uzodoba ngaphansi kwezimo ezithengisa izakhamuzi zaseNingizimu Afrika.