Effective Transfer Learning Between Morphologically Similar Languages: A Case Study on English-Zulu Translation

Samantha Ball
1603701

Innocent McHechesi
2497239

Muhammad Umair Nasir
2396876

Abstract

Stemming from the limited availability of datasets and textual resources for low-resource languages such as isiZulu, there is a significant need to be able to harness knowledge from pre-trained models to improve low resource machine translation. Moreover, a lack of techniques to handle the complexities of morphologically rich languages has compounded the unequal development of translation models, with many widely spoken African languages being left behind. This study explores the potential benefits of transfer learning and different tokenization schemes in an English-Zulu translation framework. The results indicate the value of transfer learning from closely related languages to enhance performance of low-resource translation models, thus providing a key strategy for low-resource translation going forward.

1 Introduction

Neural machine translation aims to automate the translation of text or speech from one language to another utilising neural networks (Nyoni and Bassett, 2021). Consequently, the performance of neural machine translation (NMT) models is highly dependent on the availability of large parallel corpora to provide sufficient training data. Low-resource languages which are under-represented in internet sources lack suitable training corpora and therefore suffer from limited development, obtaining poor translation performance. This phenomenon is exacerbated by a lack of content creators, dataset curators and language specialists, resulting in barriers at many stages in the translation process (Lakew et al., 2020).

Therefore, due to the historical focus on dominant languages such as English in the development of neural machine translation (NMT) models, low-resource and morphologically complex languages remain a challenge for current translation systems. Due to limited resources in terms of both computational expense and available datasets, it is vital to be able leverage knowledge from current pretrained models to provide more effective solutions. Moreover, due to the morphologically rich nature of many low-resource languages, default tokenization schemes are inadequate, necessitating the use of more appropriate techniques. In this investigation, the effects of transfer learning from closely related languages, as well as different tokenization schemes, is explored in the context of English to Zulu translation.

2 Background

Previous studies have indicated poor translation performance for the isiZulu languages due to its morphological complexity and limited available data (Martinus and Abbott, 2019). The challenging nature of English-isiZulu translation is highlighted in a benchmark of five, low-resource African languages by Martinus and Abbott (2019), where isiZulu obtains a much poorer BLEU score in comparison to other evaluated languages. The study suggests that the collection of higher quality datasets for isiZulu would greatly benefit translation performance.

Furthermore, the challenges associated with the morphological complexity of Nguni languages such as isiZulu are tackled in a study by Moeng et al. (2021). The investigation explores the use of supervised sequence-to-sequence models to tokenize isiZulu, isiXhosa, isiNdebele and siSwati sentences, demonstrating promising results for improved segmentation of morphologically

complex Nguni languages.

A notable study by Nyoni and Bassett (2021) compares the use of zero-shot learning, transfer learning and multi-lingual learning on three Bantu languages, namely isiZulu, isiXhosa and Shona. The results indicate that multi-lingual learning where a many-to-many model was trained using three different language pairs, English-isiZulu, English-isiXhosa and isiXhosa-isiZulu led to optimal results on their custom dataset.

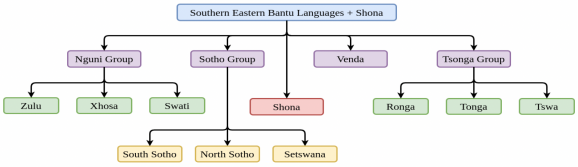

Figure 1: Southern Bantu language family tree (Nyoni and Bassett, 2021).

In addition, the study found that transfer learning from a closely related Bantu language is highly effective for low resource translation models, with statistically significant results being obtained when transfer learning to isiZulu using the pretrained English-to-isiXhosa model (Nyoni and Bassett, 2021). In contrast, transfer learning from the English-to-Shona model did not yield any statistically significant improvement, indicating the role of morphological similarity in the transfer learning process.

Similarly, this study aims to investigate whether transfer learning from a morphologically similar language will be effective on the novel, high-quality Umsuka English-isiZulu parallel corpus, and whether the tokenization scheme utilised will impact translation performance.

3 Methodology

This investigation evaluates several models pretrained on different language pairs on a recently release English-Zulu parallel corpus. In addition, three different tokenization schemes are explored to ascertain their effectiveness for the morphologically complex isiZulu language. The dataset utilised to fine-tune and benchmark the models is discussed below.

3.1 Dataset

The Umsuka English-isiZulu Parallel Corpus (Mabuya et al., 2021) provides a novel, high-quality parallel dataset for machine translation, containing English sentences sampled from both News Crawl datasets which were then translated into isiZulu, and isiZulu sentences from the NCHLT monolingual corpus and UKZN isiZulu National monolingual corpus, which were then translated into English. Each translation was performed twice, by two differing translators, due to the high morphological complexity of the isiZulu language. The dataset is open source and available from the Zenodo platform[1].

3.2 Tokenization Schemes

Due to the nature of isiZulu as an agglutinative language, simple tokenization schemes relying on the separation of sentences into words as the lowest unit of meaning will result in very poor performance. Therefore three different tokenization strategies are explored, namely *Byte Pair Encoding*, *WordPiece* and *SentencePiece*.

(i) Byte Pair Encoding (BPE)

Byte Pair Encoding (BPE) (Sennrich et al., 2015) provides a means of sub-word tokenization which is more appropriate for morphologically complex languages. Many low-resourced languages are agglutinative and therefore each word is made up of several parts which each encode a different concept or piece of information. Therefore tokenizers that rely on separating on white space and consider each word as the smallest unit are unsuitable for many low-resourced languages as they will not correctly separate the units of meaning. In contrast, Byte Pair Encoding provides a more suitable solution since it can tokenize into subwords and therefore into units of meaning that are smaller than words.

In addition, word-level tokenization can only process a fixed number of words and therefore has limited or fixed vocabulary. This was cause the tokenization to perform poorly for unknown words. In contrast, subword-level tokenization allows for a more adaptable vocabulary as unknown words can be encoded as sequences of subwords and relationships between subwords can be inferred. This is particularly useful for low resource languages as there may be a larger proportion of unknown and

[1] https://zenodo.org/record/5035171#.YZvn1fFBy3J

rare words since there is less available training data.

Several limitations of BPE include increased computational expense, the challenge of finding an optimal vocabulary size for the given task as well as ambiguity stemming from the fact that the same input can be represented by different encodings.

(ii) WordPiece

WordPiece is an alternative sub-word tokenization scheme proposed by Schuster et al. (2012). Similarly to BPE, the vocabulary is initialised using all the characters and then the most frequent combinations of characters are iteratively added to the vocabulary. While in BPE, the next word unit added to the vocbulary is the pair which has the highest combined frequency, in WordPiece the next word unit is selected as the symbol pair which maximises the likelihood of the data (Schuster and Nakajima, 2012). In addition to this bottom-up approach, there is also a top-down version of the WordPiece tokenization scheme. Therefore the major contrast between BPE and WordPiece is the way in which new word units or symbol pairs are added to the vocabulary, with WordPiece taking a more probabilistic approach.

(iii) SentencePiece

Lastly, SentencePiece is a subword tokenizor which offers the advantages of speed and subword regularization. SentencePiece is very efficient as it is implemented in C++. In addition, SentencePiece works on languages with and without whitespace. SentencePiece utilises either BPE or the Unigram algorithm to construct the underlying vocabulary. In this investigation, the Unigram algorithm as proposed by Kudo et al. (2018) was utilised in conjunction with SentencePiece.

The Unigram approach to building the vocabulary differs from BPE and WordPiece since it initialises a large number of symbols as the base vocabulary and then successively removes symbols according to a defined loss function. Symbols are removed based on how much the overall loss would increase if a given symbol was removed (Kudo, 2018). The percentage of symbols that contribute the lowest increase in loss are then removed from the vocabulary. While BPE and WordPiece both represent a greedy approach to tokenization, the Unigram algorithm is a fully probabilistic method which aims to predict which of the possible encodings is most appropriate, therefore tackling the issue of ambiguity present in BPE.

3.3 Models

The three models tested are based on the MarianMT model (Junczys-Dowmunt et al., 2018) which is constructed using a Transformer architecture. Each model is pretrained on a different set of language pairs from the Helsinki Corpus.

MarianMT

MarianMT (Junczys-Dowmunt et al., 2018) is a toolkit for neural machine translation written in C++ with over 1000 models trained on different language pairs available from the HuggingFace library. Each model is based on a Transformer encoder-decoder structure with 6 layers in each component (Junczys-Dowmunt et al., 2018). From the available models, three pre-trained models were selected, representing pre-training on a closely related language, pre-training on a more distantly related language within the same family and pre-training on multiple unrelated languages respectively. Since each model was based on the same architecture, this allowed for a controlled comparison of the language pairs used for pre-training, as well as the effects of various tokenization schemes, as any discrepancies due to architectural differences were discounted.

(a) English-Xhosa

Since isiXhosa and isiZulu are both part of the Nguni branch of Bantu languages, isiXhosa is closely related to isiZulu in the Bantu language family tree (Nyoni and Bassett, 2021). Therefore the MarianMT model pre-trained on English-Xhosa pairs is selected for fine-tuning on the Umsuka English-isiZulu parallel corpus.

(b) English-Swahili

Secondly, another Bantu language, Kiswahili was explored to determine the effects of transfer learning from another language within the Bantu family which is not as closely related to the target isiZulu language. While isiZulu is classified as a

Southern Bantu and Nguni language, Kiswahili is part of the Northeast Bantu and Sabaki languages (Nurse et al., 1993).

(c) Multilingual-Romance

Lastly, the impact of pre-training on multiple unrelated languages was investigated through fine-tuning and evaluating a MarianMT model pre-trained on 48 Romance languages, including French, Spanish and Italian. Since the Romance languages differ strongly from the morphology of the Bantu languages, this multilingual model provides a good contrast to the other models considered.

4 Results

Each model was benchmarked on the evaluation set using the BLEU score as tabulated in Table 1 below. It can be observed that the optimal model is given by the MarianMT model pre-trained on the English-Xhosa dataset with SentencePiece tokenization. This confirms our expectation that transfer learning from a morphologically similar language would result in improved performance.

In Fig. 2 below, we can observe that the MarianMT model pre-trained on the English-Xhosa dataset outperforms both the English-Swahili and Multi-lingual Romance models, obtaining a final BLEU score of 13.387. This result suggests that the morphological similarities between the isiZulu and isiXhosa languages plays a strong role in the benefits attained through fine-tuning.

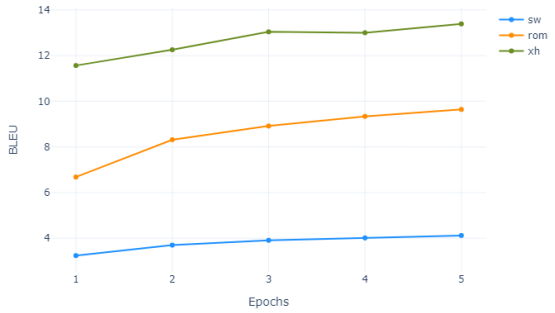

Figure 2: BLEU scores per epoch according to different pre-training languages, using the best-performing SentencePiece tokenization, indicating much better results for pre-training on the English-Xhosa dataset.

Notably, although Kiswahili is within the Bantu language family, the model pre-trained on the English-Swahili dataset obtains the poorest results. This suggests that related languages must be carefully selected for pre-training since selecting languages from the same overall family does not necessarily guarantee better results.

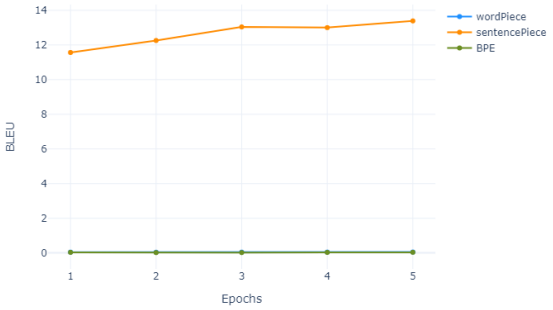

Figure 3: BLEU scores per epoch according to tokenization type, with En-Xh pre-training, indicating much better results for SentencePiece tokenization.

In addition, we observe from Table I that the SentencePiece tokenization scheme performed considerably better than the BPE and WordPiece tokenization schemes across all models. This result is visualised in Fig. 3, indicating the large discrepancy in BLEU score obtained using SentencePiece tokenization in comparison to BPE and WordPiece, with pre-training kept constant. This is likely due to the mechanism of the underlying Unigram algorithm which represents a fully probabilistic approach where symbols are removed from a large base vocabulary based on a loss function, and therefore the most likely symbols are retained.

Following identification of the optimal model and tokenization scheme, the MarianMT model pre-trained on the En-Xh dataset with SentencePiece tokenization was fine-tuned for 50 epochs, giving a final optimal BLEU score of 14.461.

5 Analysis

We now present an analysis of the results in light of both the underlying theory and previous literature. In order to further understand the effects of pre-training on different languages, the datasets used for pre-training of the MarianMT models were inspected. Notably, although the number of tokens in the training dataset is much

Table 1: Quantitative metrics indicating performance of each NMT model on the Umsuka English-isiZulu Parallel Corpus.

Model	Pretraining	Tokenization	BLEU Score
MarianMT	English-Xhosa	BPE	0.035
		WordPiece	0.051
		SentencePiece	13.387
MarianMT	English-Swahili	BPE	0.027
		WordPiece	0.239
		SentencePiece	4.118
MarianMT	Multilingual-Romance	BPE	0.053
		WordPiece	0.528
		SentencePiece	9.640

greater for the English-Swahili dataset than for the English-Xhosa dataset, the model pretrained on the English-Xhosa dataset still outperforms that of the Swahili dataset. This further underlines the value of identifying closely related languages for pre-training. This suggests that while the quantity of training data matters, the morphological similarity is still more important when transfer learning.

Lastly, a significant result is obtained through the continued fine-tuning of the optimal model, resulting in a final BLEU score of 14.461. This score is comparable to that of previous literature focusing on English-isiZulu translation. However, in order to fully compare against other models, the respective models would need undergo evaluation on the same benchmark. An investigation of this nature is suggested for further analysis.

6 Impact Statement

The potential impacts of this investigation can be explored in light of the possible contributions, risks and societal impact.

6.1 Applications and Benefits

The study poses potential benefits to further research into low-resource languages as it motivates careful choice of the pre-trained model used for transfer learning in order to improve performance on low resource languages. This could provide a vital tool to improve the efficiency and performance of low resource translation pipelines, especially in resource-constrained environments. In addition, this principle could be applied more broadly to other language groups with morphologically similar languages.

Moreover, effective transfer learning provides the additional advantage of promoting decreased computational expense since prior knowledge from previously trained networks can be leveraged effectively. This could work to mitigate the substantial detrimental environmental impact stemming from the intensive GPU training required to train neural machine translation models. This is critical to ensure sustainable development of machine translation models by minimising resource waste.

6.2 Limitations and Drawbacks

It should be noted that any conclusions drawn from the study are based on the BLEU score as the sole evaluation metric. This may provide a limited view of the true translation performance as it is based on n-gram similarity and does not necessarily measure whether the meaning of a sentence has been captured. A further improvement could be to conduct a similar study with additional expertise from a linguistic specialist to verify whether the output of the translation models is valid.

Lastly, although significant improvements are obtained through the use of transfer learning from the morphologically similar isiXhosa language, the overall BLEU scores obtained for the English to isiZulu translation models remain much lower than those for high-resource languages. This suggests that further development is necessary to develop a robust model for English-Zulu translation.

6.3 Social Impact

Societal impacts of low resource neural machine translation include furthering accessibility of information to under-represented languages and working to close the digital divide between high-resource and low-resource languages. Machine translation is an essential component of applications ranging from voice-assisted smart-phone ap-

plications that provide healthcare to rural communities to ensuring multi-lingual access to educational materials. Therefore it is vital that machine translation technology is accessible and functional for low-resource languages to be able to build valuable tools which could have a beneficial societal impact.

7 Conclusion

English-isiZulu translation has historically obtained poor results on translation benchmarks due to a lack of high-quality training data and appropriate tokenization schemes able to handle the agglutinative structure of isiZulu sentences. In this investigation, the challenges of isiZulu translation in terms of both morphological complexity and a lack of textual resources are explored using the recently released Umsuka English-isiZulu Parallel Corpus. In order to investigate the effects of different subword tokenization schemes as well as the impact of the pre-trained model selected for transfer learning, several models were fine-tuned and benchmarked on the Umsuka dataset.

Specifically, the performance of the BPE, WordPiece and SentencePiece subword tokenization schemes were compared using existing MariantMT models pre-trained on English-Xhosa, English-Swahili and English-Multilingual Romance languages respectively. The study found that SentencePiece performed considerably better for all tested models, while the pre-trained English-Xhosa model attained the optimal results. Thus, the results indicate that transfer learning is particularly effective when languages are within the same sub-family while transfer learning is less effective when the model is pre-trained on a more distantly related language.

Therefore, this study motivates careful choice of the pre-trained model used for transfer learning, utilising existing knowledge of language family trees, to promote improved performance of low resource translation. In addition, we have open-sourced[2] our best model which was fine-tuned for 50 epochs using the original MarianMT model pre-trained on the English-Xhosa language pair with SentencePiece tokenization, obtaining a final BLEU score of 14.461.

[2]Available online at https://huggingface.co/MUNasir/opus-en-xh-umsuka-en-zu

References

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *arXiv preprint arXiv:2003.14402*.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka english - isizulu parallel corpus.

Laura Martinus and Jade Z Abbott. 2019. A focus on neural machine translation for african languages. *arXiv preprint arXiv:1906.05685*.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and surface morphological segmentation for nguni languages. *arXiv preprint arXiv:2104.00767*.

Derek Nurse, Thomas J Hinnebusch, and Gérard Philipson. 1993. *Swahili and Sabaki: A linguistic history*, volume 121. Univ of California Press.

Evander Nyoni and Bruce A Bassett. 2021. Low-resource neural machine translation for southern african languages. *arXiv preprint arXiv:2104.00366*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.