Geographical Distance Is The New Hyperparameter: A Case Study Of Finding The Optimal Pre-trained Language For English-isiZulu Machine Translation.

Muhammad Umair Nasir[1], **Innocent Amos Mchechesi**[2]

[1] Ominor AI, [2] University of the Witwatersrand, South Africa

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 in an English-isiZulu translation framework. The results indicate the value of transfer learning from closely related languages to enhance the performance of low-resource translation models, thus providing a key strategy for low-resource translation going forward. We gathered results from 8 different language corpora, including one multilingual corpus, and saw that isiXhosa-isiZulu outperformed all languages, with a BLEU score of 8.56 on the test set which was better from the multi-lingual corpora pre-trained model by 2.73. We also derived a new coefficient, *Nasir's Geographical Distance Coefficient (NGDC)* which provides an easy selection of languages for the pre-trained models. NGDC also indicated that isiXhosa should be selected as the language for the pre-trained model.

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; Zoph et al., 2016; Sennrich and Zhang, 2019).

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 (Haddow et al., 2021; Koehn and Knowles, 2017). Due to limited resources in terms of both computational expense and available datasets, it is vital to be able to leverage knowledge from current pre-trained models to provide more effective solutions. Therefore, in this investigation, the effects of transfer learning from closely related languages, as well as comparison with high-resourced languages for pre-trained scenario, is explored in the context of English to Zulu translation.

Furthermore, this study derives the Nasir's Geographical Distance coefficient. *Geographical Distance (GD)* (Holman et al., 2007) has been studied for various scientific research areas (Bei et al., 2021; Krajsa and Fojtova, 2011; Riginos and Nachman, 2001) as it provides deep insights in many aspects. We will also use GD as a hyperparameter for an attempt to get a language for a pre-trained model in an effective and with a $O(n)$ complexity. Although there are many ways to find GD, we will use literal approximation of distance in kilometers and suggest the techniques in future directions.

1.1 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

Proceedings of the Workshop on Multilingual Information Access (MIA), pages 1 - 8

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 chiShona. 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.

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.

There has been a lot of work in providing assistance to low-resourced languages for machine translation focus of the area. Neubig and Hu (2018) trained multilingual models as seed models and then continued training on low-resourced language. Sennrich et al. (2015) looks into training monolingual data with automatic back-translation (Edunov et al., 2018; Caswell et al., 2019; Edunov et al., 2019) to improve scores through only a monolingual data. Another work that utilizes back-translation for effecctive NMT training is done by Dou et al. (2020). Koneru et al. (2022) proposes a cost-effective training procedure to increase the performance of models on NMT tasks, utilizing a small number of annotated sentences and dictionary entries. Park et al. (2020) looked into decoding strategies for low-resourced languages in an attempt to improve training. Nguyen and Chiang (2017) looked into related languages to a target language for low-resourced languages to prove effectiveness of similar languages.

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 if so, how does it perform when we use high-resourced mono- and multi-lingual corpora. This study will also derive a formula which will ease the way for selecting a language for a pre-trained model.

2 Methodology

This investigation evaluates several models pre-trained on different language pairs, both low- and high-resourced, on a recently release English-Zulu parallel corpus. The dataset utilized to fine-tune and benchmark the models is discussed below.

2.1 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. This also serves the purpose of considering one translation as a reference and the other as target. This can be validated as both have been translated by human annotators and are different from each other. The dataset is publicly available from the Zenodo platform[1].

2.2 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 (Junczys-Dowmunt et al., 2018) is a toolkit for neural machine translation written in C++ with over 1000 models trained on different language pairs from OPUS[2], available at the Hug-

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

[2] https://opus.nlpl.eu/

gingFace Model Hub[3]. 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, 8 pre-trained models were selected[4], 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, with less and more data, 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 any discrepancies due to architectural differences were discounted.

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). As well as Shona, or chiShona, is selected as it is also a part of Southern Bantu language group (Nyoni and Bassett, 2021). 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).

Twi, or Akan-kasa, is spoken in Ghana, has been selected to have a representation from Western Africa and to explore the effects a dialect of the Akan language on fine-tuning isiZulu. Luganda is selected as a representation from Niger-Congo family of languages and is spoken in East-African Country of Uganda. This will able us to explore the fine-tuning regime in Niger-Congo languages.

Arabic and French are selected as they are morphologically very different and are considered to be high-resourced (Ali et al., 2014; Besacier et al., 2014). We explore effects of fine-tuning high-resourced languages with different morphologies. As the notion of having more and multi-lingual data will be better for fine-tuning, we select a corpus of Romance languages, which is created by joining 48 Romance languages including French, Italian, Spanish, Walloon, Catalan, Occitan, Romansh etc. We include Romance languages so that we can cover the aspect of big multi-lingual corpora being fine-tuned on low-resourced isiZulu and to prove our hypothesis.

[3] https://huggingface.co/

[4] https://github.com/umair-nasir14/NGDC

2.3 Implementation Reproducibility

We believe all experiments must be *Reproducible*. To achieve this we are open-sourcing our code on GitHub (added in the footnote previously).

3 Results

Each model was benchmarked on the test set using the BLEU(Papineni et al., 2002) 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. This confirms our hypothesis that transfer learning from a geographically distant language would result in poor performance. Here GD is in Kilometers (Km) and corpus size is in Number Of Sentences in millions (M).

In Fig. 1 below, we can observe that the MarianMT model pre-trained on the English-Xhosa dataset outperforms all other models by a good margin, obtaining a final BLEU score of 8.56. This result suggests that the morphological similarities between the isiZulu and isiXhosa languages plays a strong role in the benefits attained through fine-tuning.

Following identification of the optimal model, the MarianMT model pre-trained on the En-Xh dataset was further fine-tuned for 75 epochs on Umsuka dataset, giving a final optimal BLEU score of 17.61 on training set and 13.73 on test.

4 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 sentences in English-Xhosa dataset is in order of magnitudes less than Romance languages corpus but still performs better. This justifies our hypothesis and opens up a path to effective fine-tuning through the knowladge of morphologies and not by adding multiple languages into a single corpus. Arabic and French having approximately 5 and 23 times more data also suggests the above mentioned hypothesis that with closer GD and lesser data is much better, in many ways, than larger data and farther GD.

Other Bantu languages that were selected, Kiswahili and chiShona performed almost similar to Arabic and French with order of magnitudes of lesser data which suggests that even if they are

Language(s)	BLEU(Val)	BLEU(Test)	Corpus Size(NOS)	GD(KM)
isiXhosa	10.20	8.56	20.7	1000
Romance	7.76	5.83	1232.7	13094.4
Arabic	5.76	3.07	102.8	5205
French	5.42	3.91	479.1	13094
Kiswahili	5.28	3.97	9.1	3783.1
chiShona	4.32	2.83	0.1	1584
Twi	1.91	1.34	0.047	7962
Luganda	0.94	0.55	0.039	4883.7

Table 1: BLEU scores, GD and corpora size

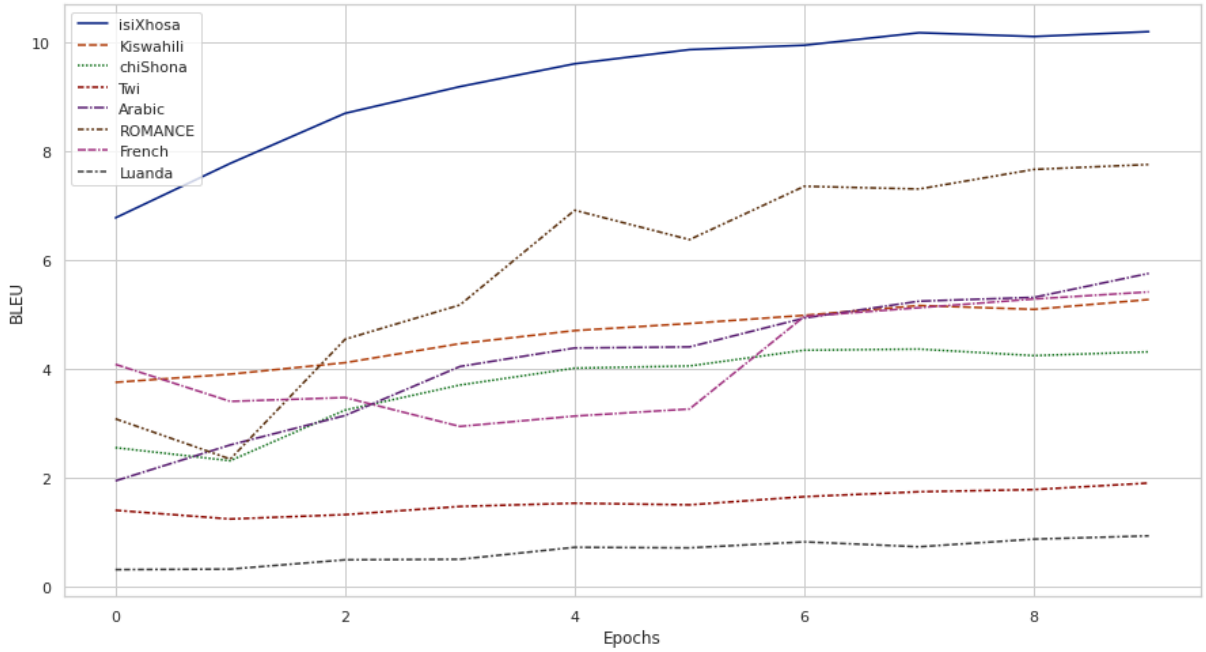

Figure 1: BLEU scores per epoch according to different pre-training languages, indicates high performance of morphologically similar isiXhosa, which outperforms a model trained on a very large corpora and rest of corpora.

not as similar to isiZulu, the distance being very close to where isiZulu is spoken tends to have a great impact. We speak similar languages in neighbouring cities and countries which should have an effect on the model and so the result suggests. Twi and Luganda, having very less data and higher GD, gives us very poor results.

From Table 1, we also observe that distance between the target language and the language from a pre-trained model is a very important factor. Alone, to a good extent it can serve the purpose of choosing the language of pre-trained model but we want to look one step deeper as one can argue that Romance languages corpora, French and Arabic perform relatively better but the distances are larger. Thus we also look into Size of Corpus (Table 1). Which forces us to think about deriving a relationship that involves both distance and the size. This will be explained in the upcoming sub-section.

4.1 Nasir's Geographical Distance Coefficient

In Figure 2 we can observe that there is a sensible relationship between BLEU scores and distance, and as a rule of thumb there should always be a relationship with corpus size (Lin et al., 2019). With further analysis we can deduce that neither distance alone nor corpus size alone can be taken for granted when selecting a language for pre-trained model. Thus, we derive a formula which takes into account both distance and corpus size in account. This formula is intended to be used before training to know which language corpora to select.

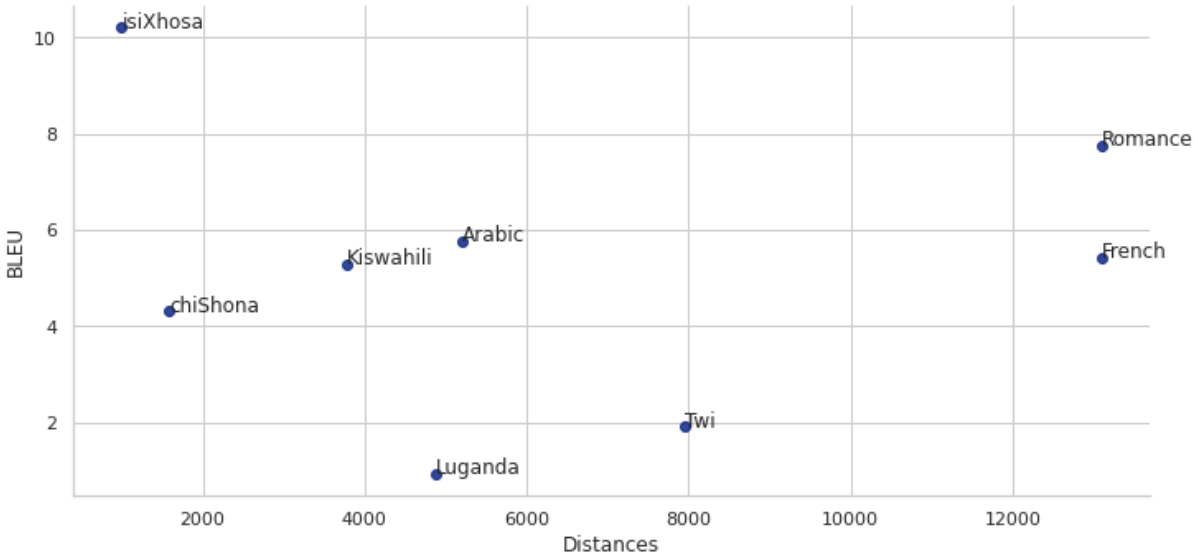

Figure 2: Relationship between BLEU scores and distance (KM) of places where languages are spoken from the place where isiZulu is spoken.

$$z = \frac{cD}{(1-c)S}$$

$$\delta = \begin{cases} 1, & \text{if } D \geq D_{max} \\ \frac{exp(z)}{1+exp(z)}, & \text{otherwise} \end{cases}$$

where D is the distance between language to fine-tune and language of the pre-trained model, S is the size of corpus, c is the weight coefficient, set to 0.4, which could act as hyperparameter. D_{max} is also hyperparameter to be tuned when it is being used in different languages in different parts of the world. δ is the coefficient we are introducing, *Nasir's Geographical Distance Coefficient (NGDC)*. The goal here is to minimize NGDC.

Table 2, Figures 3 and 4 shows the results and effectiveness of our introduced NPC. We can observe that without imposing penalty we have Romance languages, Arabic and French as desired pre-trained model languages along with isiXhosa and Kiswahili, which makes absolute sense as some have more data and others are near to target language but we want to have morphologically closer languages which will get better results. It would also be better if lesser carbon footprint is left and lesser training resources are used. Thus, with the penalty we only get isiXhosa and Kiswahili as desired ones, which will eventually be better in all perspectives.

5 Impact Statement

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

5.1 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.

5.2 Limitations and Drawbacks

It should be noted that any conclusions drawn from the study are based on the BLEU score as the sole

Language(s)	BLEU	NGDC(With Penalty)	NGDC(Without Penalty)
isiXhosa	10.20	0.5080	0.5080
Romance	7.76	1.0000	0.5007
Arabic	5.76	1.0000	0.5084
French	5.42	1.0000	0.5045
Kiswahili	5.28	0.5688	0.5688
chiShona	4.32	0.9999	0.9999
Twi	1.91	1.0000	1.0000
Luganda	0.94	1.0000	1.0000

Table 2: NGDC with and without Penalty.

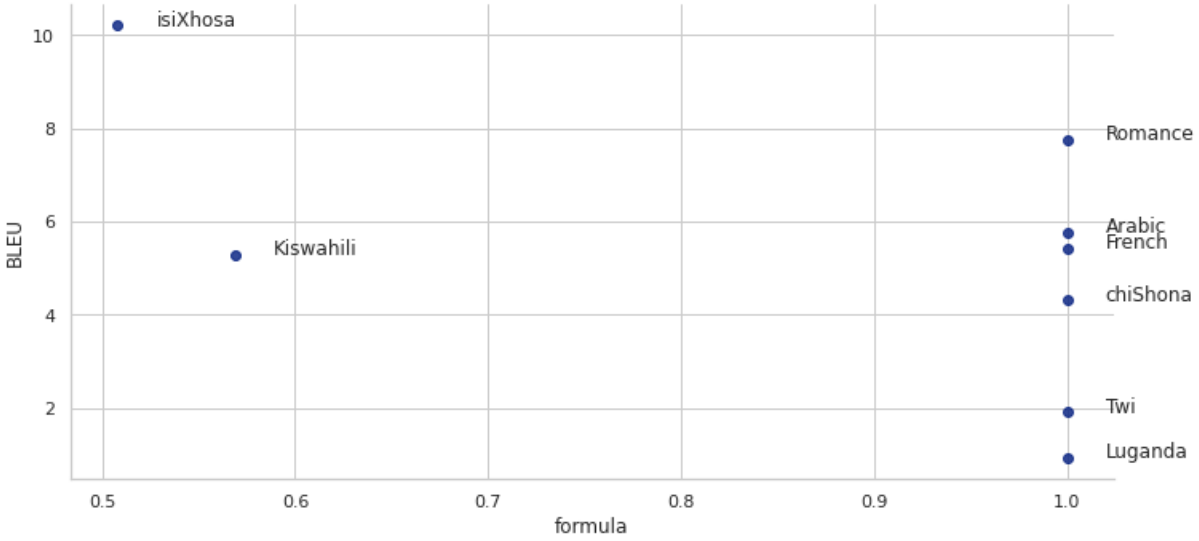

Figure 3: NGDC with Penalty

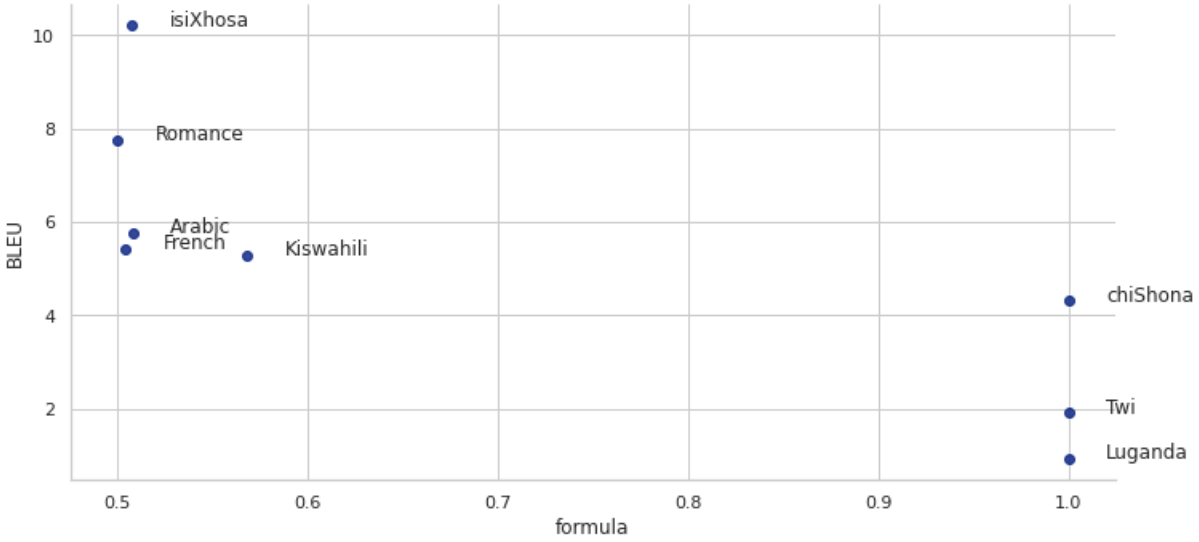

Figure 4: NGDC without Penalty

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.

5.3 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 applications 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.

6 Conclusion and Future Directions

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 the impact of the pre-trained model selected for transfer learning, several models were fine-tuned and benchmarked on the Umsuka dataset.

MariantMT models pre-trained on English-Xhosa, English-Swahili, English-Shona, English-Twi, English-Luganda, English-Arabic, English-French and English-Multilingual Romance languages, respectively. The study found that the pre-trained English-Xhosa model attained the optimal results with a handsome margin. 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, no matter the size of the data to an extreme extent. We have also introduced a novel *Nasir's Geographical Distance Coefficient* which will help researchers find a language for pre-trained model effectively and will result in using less resources.

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[5] our best model which was fine-tuned for 75 epochs using the original MarianMT model pre-trained on the English-Xhosa language pair, obtaining a final BLEU score of 17.61 on train while 13.73 on test set. We have also gathered all model cards for the models that were used for further experimentation.

This study yeilds promising future directions as the experiment was done on only 8 corpora. We suggest to increase the number and observe the derivation of the result. We also suggest to combine Bantu language as one multi-lingual corpora and observe the result. The experiment has been done on a novel Umsuka parallel corpora, the study should extend to more common benchmarks. This study should extend to different low-resourced languages of different continents of our world. We have derived a formula that takes into the account just the distance and the size of corpora, a promising research would be to derive a formula that takes morphologies and/or phonologies and fives a distance based on that. With NGDC at hand, it motivates to create a framework where one enters a target language, a D_{max} and a value for weight coefficient c and gets desirable models to train on. There are many precise ways of finding GD, such as Lambert's formula (Lambert, 1942) and Vincenty's formula (Vincenty, 1975) which may enhance NGDC's performance. It also opens up ways to introduce morphology in the formula, which we expect it to improve the overall selection of the models.

References

Ahmed Ali, Hamdy Mubarak, and Stephan Vogel. 2014. Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Papers*.

Eva Bei, Mikołaj Zarzycki, Val Morrison, and Noa Vilchinsky. 2021. Motivations and willingness to provide care from a geographical distance, and the impact of distance care on caregivers' mental and physical health: A mixed-method systematic review protocol. *BMJ open*, 11(7):e045660.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Introduction to the special issue on processing under-resourced languages.

[5] https://huggingface.co/MUNasir/umsuka-en-zu

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2021. Survey of low-resource machine translation. *arXiv preprint arXiv:2109.00486*.

Eric W Holman, Christian Schulze, Dietrich Stauffer, and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations.

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*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Sai Koneru, Danni Liu, and Jan Niehues. 2022. Cost-effective training in low-resource neural machine translation. *arXiv preprint arXiv:2201.05700*.

Ondrej Krajsa and Lucie Fojtova. 2011. Rtt measurement and its dependence on the real geographical distance. In *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*, pages 231–234. IEEE.

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*.

Walter D Lambert. 1942. The distance between two widely separated points on the surface of the earth. *Journal of the Washington Academy of Sciences*, 32(5):125–130.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*.

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*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chanjun Park, Yeongwook Yang, Kinam Park, and Heuiseok Lim. 2020. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

C Riginos and MW Nachman. 2001. Population subdivision in marine environments: the contributions of biogeography, geographical distance and discontinuous habitat to genetic differentiation in a blennioid fish, axoclinus nigricaudus. *Molecular ecology*, 10(6):1439–1453.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. *arXiv preprint arXiv:1905.11901*.

Thaddeus Vincenty. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.