UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Final Paper 2021

A Comparison of Data Augmentation Techniques for Nguni Language Statistical and Neural Machine Translation models

Fezeka Nzama

Project Abbreviation: SMT-NMT

Supervisor(s): Jan Buys

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	0
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	10
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	10
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	0
Total marks		**80**	**80**

A Comparison of Data Augmentation Techniques for Nguni Language Statistical Machine Translation models

Fezeka Nzama
Department Of Computer Science
University Of Cape Town
Cape Town, South Africa
nzmfez001@myuct.ac.za

Abstract

The advent of the internet, and rise in available written text online has lead to significant improvements in the field of machine translation. These improvements have been particularly evident in the high resource language context where there exists large volumes of parallel corpora for languages. However, this has not been in the case for low resource languages such as the Nguni language group of South Africa as there is minimal existing parallel corpora for these languages. Given that the most important factor to high quality machine translation is the existence of large volumes of training data a number of data augmentation techniques have been suggested. This paper provides a comparison two of theses data augmentation techniques, namely back-translation data augmentation and multilingual data augmentation, against a baseline system for the Nguni languages of isiZulu and isiXhosa. The expectation was that in both cases the back-translation system and multilingual system would outperform the baseline system with the multilingual system providing the best performance. The results, however, showed that in both the isiZulu and isiXhosa context the back-translation system provided the best performance.The baseline system outperformed the multilingual system in the isiZulu context and the multilingual system outperformed the baseline system in the isiXhosa context. These results are further explored and further research avenues suggested.

1 Introduction

In an increasingly intertwined global world, machine translation offers a means to facilitate interlingual communication and thus economic and educational inclusion. Machine translation is the use of computer systems to perform text-based translations from some source language to some target language [13]. The current state of the art machine translation method is neural machine translation (NMT), which makes use of the neural network architecture to perform translations [14, 26]. The predecessor to NMT is statistical machine translation (SMT) which is based on Bayes Theorem for conditional probability, and continues to have potential applications in the low resource language context [8**?** , 9].

The advent of NMT has resulted in significant improvement in translation quality for high resource languages, due to the large existing bodies of parallel corpora for these languages [8, 11, 15, 19, 22]. However, due to the large amounts of data required to optimise NMT, the improvements witnessed in the high resource context have not been observed for translations for low resource languages. These low resources such as the Nguni languages of South Africa, have limited existing bodies of parallel corpora. In the low resource setting, SMT has generally been shown to outperform NMT [3, 15, 23] despite more recent studies in which NMT has outperformed SMT for short sentences or when very finely tuned NMT systems have been used for particular contexts [3, 15].

As previously alluded to, the primary challenge facing the low resource machine translation is one of data sparsity. Whilst numerous data augmentation techniques have been suggested in previous studies, [2, 4, 9, 10, 24] few of these techniques have been used in the Nguni context and there currently exist no comparison of these data augmentation techniques in the Nguni language context. As such this study aims to compare baseline SMT systems for the Nguni languages of isiZulu and isiXhosa with SMT systems that make use of data augmented with the use of backtranslation and multilingual training data for each language. It is hypothesised that the models that make use of augmented data will outperform the baseline SMT system, with the model making use of a multilingual training dataset yielding the best performance.

The remainder of this paper is organised as follows: Section 2 will provide an outline of the background information on SMT, highlighting the two main components of an SMT system and review previous work done for low resource SMR. Section 3 will provide a description of the data and data preprocessing steps used in this study, outline the tools language and translation model tools used and the describe the different types of SMT systems used in this study. Section 4 will detail the evaluation process and data used for this study, with the results of this study presented and discussed in Section 5. Finally, in section 6 the conclusion derived from this study will be presented with a discussion on further work that may be conducted as a result.

2 Background & Related Work

The most widely used form of SMT is phrase-based SMT, which seeks to perform machine translation at phrase level. This translation model is based on the Bayes Theorem for conditional probability Pr(T|S) where T is the target language and S is the source language [21]. This conditional probability is expressed as follows:

$$Pr(T|S) = \frac{Pr(T)Pr(S|T)}{Pr(S)} \tag{1}$$

SMT systems consist of 2 main components, the language model and the translation model. This language model is represented in Bayes Theorem as Pr(T) whereas the translation model is expressed as

Pr(S|T). This formula aims to maximise Pr(T|S) and can be simplified as follows:

$$T' = argmax_T[Pr(T) * Pr(S|T)] \tag{2}$$

Combined the 2 elements of the SMT model as shown above are referred to as the noisy-channel model [13].

2.1 Language Modelling

The language model works to capture information about word ordering within a target language sentence, and thus allows for greater fluency for translations [13]. The language model requires only a set of substantial monolingual data in the target language.

N-Grams The most commonly used method for deriving this language model is the Back-Off N-gram model [26]. This model seeks to estimate the probability of a word appearing at the end of a sequence given a list of n-1 words have already been seen [13, 21] such that $P(w|h)$ is the probability of seeing some word w given a history of h words have already been observed. For example given a trigram model is being used, and a history h of *"the boy"*, the probability of seeing the word *"goes"* can be represented by the following formula:

$$P(goes|the\ boy) = \frac{C(the\ boy\ goes)}{C(the\ boy)} \tag{3}$$

where the probability is given by the relative frequency counts of seeing the sentence phrase *"the boy"* followed by the word *"goes"* [13, 18].

N-grams are be highly effective in the high resource language setting, whilst also being simplistic to use. They do, however, suffer from data sparsity issues in the low resource setting, leading to translation fluency inconsistencies [13, 21]. To overcome this issue N-grams make use of smoothing techniques to account for known words ie. words that are recognised as part of the vocabulary of the language appearing in unseen contexts. This is necessary to ensure that a probability of 0 is not given to these unseen contexts [18]. A number of methods maybe used to perform smoothing are outlined below.

- Backoff: This smoothing technique makes use of estimations for higher order n-grams based on lower order n-grams. For example, give a trigram with a probability of $P(w_n|w_{n-1}w_{n-2})$ where w_n has not been seen in the context of the trigram previously, an estimated probability is given based on the bigram $P(w_n|w_{n-1})$ [13].
- Interpolation: Simailar to the Backoff technique, Interpolation makes use of the n-gram hierarchy to perform smoothing. However, instead of using an estimation based solely on the lower order n-gram, interpolation makes use of the weighted sum of all the lower order n-grams. As such the smoothing of a trigram using Interpolation would involve taking a weighted sum of the trigram, bigram and unigram for the observed word w_n [13, 18].
- Kneser-Ney: This smoothing technique outperforms both of the previously discussed techniques by making use of principles observed from both Backoff and Interpolation smoothing. Kneser-Ney smoothin alternates the weightings given to the higher and lower-order n-gram based on whether the higher order n-gram is near zero or not. As such where a word is being observed in the context for the first time in a trigram a higher weighting is given to the bigram containing the word, whilst a lower weighting is given to the trigram, this is then alternated if a word has been observed previously in that trigram [13**?**].

2.2 Translation Modelling

The translation model seeks to correlate a phrase of m words in the source language A with a phrase of n words in the target language B, thus allowing translations to be made from A to B [13, 21]. As such the translation model requires as input parallel sentence aligned corpora in both the source and target languages.

2.2.1 Word Level Alignment & Word-Based SMT: The simplest form of translation modelling attempts to align languages at a word level mapping a word from the source language A to another word in the target language B [5, 13]. To do this a translation table is created that calculates the probability distribution across the entire corpora for the occurrence of the target word in relation with the source word, allowing for the target word which maximises this probability to be chosen as the translation of the source word [5]. This word-based translation model can be defined as follows:

$$Pr(T, a|S) = \frac{\in}{(l_T+1)^{l_s}} \prod_{j=1}^{l_s} t(S_j|T_{a(j)}) \tag{4}$$

with the first half of the formula, ie. $\frac{\in}{(l_T+1)^{l_s}}$, working as a normalisation factor ensuring that the sum of all possible translation of the word T, and alignment a sum to one. The second half of the formula provides the product of all the word-level translation probabilities. Whilst this model can be optimised, yielding more powerful word-level translation models, pales in comparison to models based on phrase level alignment.

2.2.2 Phrase level Alignment & Phrase based SMT: Phrase-based SMT makes use of phrases as the basic unit of translation , allowing for alignment to be made at a phrase level as opposed to a word level. A phrase is a group of one to many words with consistent word alignment. This means that for corresponding phrases there are no word alignments that fall outside of these phrases [13, 21]. This method leads to a decomposition of the translation as follows:

$$Pr(\overline{T}^I|\overline{S}^I) = \prod_{i=1}^{I} \phi(\overline{T_i}|\overline{S_i}) d(a_i - b_{i-1}) \tag{5}$$

The probability distribution model by the formula $\phi(\overline{T_i}|\overline{S_i})$ represents the phrase translation. The formula $d(a_i - b_{i-1})$ represents

the distance-based reordering modelling used. This model defines the differences in start position for related phrases within a source target sentence pair [13, 18].

The translation process for phrase-based SMT begins by splitting the sentence into phrases, translating each phrase and thereafter permuting these phrases into an order consistent with the target language [13, 16, 18]. Phrase pairs are extracted, and stored in a phrase translation table, with translation probability of a pair being estimated using relative frequency.

Phrase-level alignment, unlike word-level alignment allows for a many-to-many relationship to exist between languages. Additionally, phrase-level SMT is able to capture context for words better than word-level SMT. However, this method is more memory intensive than word-based SMT.

2.3 Evaluation Metrics

Translations are evaluated along two main axes: adequacy and fluency. Adequacy refers to the level at which a translation is able to capture the meaning of a source sentence. This includes the ability to accurately capture tone. Fluency refers to readability of the translation in the target language. This includes being grammatically correct, clear and natural [25]. There are two main classes of evaluators used to assess translation quality, namely human evaluation, and automatic evaluation.

2.3.1 Human Evaluation

Human evaluation provides the most accurate evaluation of translations. This method makes use of people to evaluate translations along the dimensions of adequacy and fluency by either rating translations out of some range or ranking a set of translations in order from best to worst [21, 25]. This method whilst providing the most accurate evaluation, can be time consuming and expensive.

2.3.2 Automatic Evaluation

BLEU: BiLingual Evalution Understudy or BLEU is the most widely used evaluation metric for machine translation and makes use of a comparison between the machine translation system's output and, some human generated reference translation [10]. BLEU scores are evaluated based on 3 main factors: (i)the translation length in comparison with the reference length, (ii) the words used in the translation in comparison with those used in the reference translation and (iii) the word order in the translation in comparison with the reference translation. These comparisons are done by comparing the n-grams of the translations and the reference translations [13, 21].

For a corpus, the BLEU algorithm counts the number of matching n-grams and returns as a score for the model a weighted average. Scores range from 0 for the lowest quality translation models, and 100 for highest quality or perfect translation models. Whilst BLEU is useful, it does fail to evaluate coherence in a document and does poorly when evaluating very different kinds of systems eg, Human-aided translation versus SMT [13, 21, 25].

NIST: The NIST metric, like the Bleu metric, makes use of n-gram comparisons between a reference translations and machine translations to determine the quality of a model. However, the NIST score differentiates itself from BLEU by scoring rarer segments higher weights. The aim here is to account for diversity in informational of translated texts [20].

2.4 Related Work

Whilst the clearest method for improving SMT for low resource languages is to increase the available corpora for these languages. An alternative method investigated by researchers is the use of pre- and post-processing rules to the machine translation process, which was found to increase BLEU and NIST scores for the Setswana, Sesotho and Arabic [8, 9]. Another method suggests the use of linguistic modules in conjunction with SMT to improve performance []. Another suggested approach makes use of a shared highresource pivot language. However, this approach suffers from increased probability of errors arising from multiple translations, whilst also being dependent on the existence of a large parallel corpora between the pivot language and both the source and target language [].

3 Design & Implementation

3.1 Data

3.1.1 Data Sources

- Autshumato parallel corpora: This dataset consists of data sourced by the South African Department of Arts and Culture as part of the Autshumato machine translation project initiated in 2007. This data consisted of sentence level, aligned corpora sourced from the government domain [7]. Both isiZulu-English and isiZUlu-Xhosa data was sourced from this project with 35489 isiZulu-English parallel sentences and 126708 isiXhosa-English parallel sentences.
- Mburisano parallel corpora: The Mburisano data set was created to aid in speech-to-speech COVID-19 mobile application. As such sentences from the medical domain were created and manually translated into the 11 official languages of South Africa to form this data set. This dataset consisted of 284 sentences in English, isiZulu and isiXhosa [17].
- JW300 parallel corpora: This is a corpus made up of over 300 languages with an average of 100 000 parallel sentences per language pair [1]. 866748 isiXhosa-English parallel sentences and
- MeMaT corpora: This is English-Xhosa parallel corpora collected as part pf the Medical Machine Translation project, with 379404 isiXhosa-English parallel sentences with 112593 of these sentences used for as monolingual data [12].
- NCHLT monolingual corpora: This is text collected in isiZulu for the 2014 National Centre for Human Language Technology(NCHLT) text project and consisted of 116618 isiZulu sentences [6].

3.1.2 Data Types:

The two main types of data used for this project were parallel corpora and monolingual corpora.

- Parallel Corpora: Parallel corpora is corpora existing in two or more languages as sentence align translations of each other. Parallel corpora was used to build the translation models of the SMT systems built.

Language Pair	Total No. Parallel sentences sourced
IsiZulu - English	576631
IsiXhosa - English	1260551

- Monolingual Corpora: Monolingual corpora is corpora existing only in the translation system's target language. The monolingual corpora used in this system was a combination of the target side parallel corpora sourced as well as corpora existing only in the target language. Monolingual data was used to build the language model for the target languages in the SMT systems built.

Language	No. Sentences
English	1472715
IsiZulu	693249
IsiXhosa	1373144

3.1.3 Preprocessing

- Normalisation: Using a python script, normalisation consisted of ensuring that all sentences in the corpora appeared on a separate line and all the characters in the corpora were lowercased so as to avoid the issues of data sparsity that may arise from capitalised words at the beginning of sentences. Additionally, data sourced in XML format was converted to .txt format and square brackets removed.
- Tokenisation: Once normalisation was complete, Byte -Pair encoding was used as the tokenisation algorithm for all data. This was done to account for the agglutinative nature in Nguni languages resulting in large vocabularies. As such BPE tokenisation allowed for the subword segmentation of words, aiding in the reduction of the vocabulary size. The python script used for this was based off of the Huggingface BPE tokeniser [1].
- Cleaning: Finally following tokenisation, the parallel corpora was cleaned, eliminating any sentences that were too long or empty from the corpora. The maximum sentence length allowed in the models created was 80 characters.
- Data segmentation: The SMT training process requires a large training dataset and a separate, smaller tuning set to improve the model accuracy. As such, once all the aforementioned preprocessing steps were completed the resulting parallel data was divided into a tuning set and training set, with the tuning set made up of a 1000 sentences from the parallel corpora and the remainder used for training.

3.2 Tools

All models were built and run locally on a Dell Vostro 5590 with and Intel Core i5 processor and Ubuntu operating system.

The Moses toolkit was used for all experiments conducted, with a phrase-based bidirectional reordering model conditioned on both the source and target languages.

The language model used for all systems created makes use of the KenLM language modelling toolkit. A 6-gram language model, with modified Kneser-Ney smoothing was used for all systems. The KenLM distribution used in this project came bundled with the Moses Decoder. The 6-gram configuration used was inspired by positive results from previous work conducted on language modelling for isiZulu and Sepedi [20].

The translation models used for all 3 model types were trained using GIZA++ and the grow-diag-final heuristic to perform word alignment. This heuristic starts with the intersection of two alignments and grows to include additional points of alignment.

Tuning on all 3 models was conducted using 1000 sentences of parallel corpora and the Minimum error rate training (MERT) line-search based method. This tuning is done based on the optimisation of the BLEU scores of the tuning set.

3.3 Baseline SMT model

The baseline systems created make use of the existing parallel and monolingual corpora. For each of the Nguni languages of isiZulu and isiXhosa, an English-to-Nguni translation model was created using the existing parallel corpora for the translation modelling component and the Nguni monolingual corpora for the language modelling component of the systems.

3.4 Backtranslation SMT model

The back-translation systems created were made up of two components:

(1) The Nguni to English translation component: A Nguni to English translation model was created for both isiZulu and isiXhosa individually. The model were trained to translate sentences from the Nguni language to English, making use of the English monolingual data to train the language model of the system. Once training and tuning were completed, Nguni language monolingual corpora was used as a source language set, and translated using the model into an English set of sentences. This set of parallel corpora is referred to as the synthetic parallel corpora.

[1] Available at: https://huggingface.co/transformers/fast_tokenizers.html

(2) The English to Nguni translation component: An English to Nguni translation model was created. This model made use of a combined set of parallel corpora, consisting of the original parallel corpora and the synthetic parallel corpora generated as previously outlined, as training data. The Nguni language monolingual data was used to create a 6-gram language model to be used in conjunction with the Nguni-English translation model.

3.5 Multilingual SMT model

The multilingual translation system makes use of the combined parallel corpora for both isiZulu and isiXhosa. As such in the case of the isiZulu multilingual translation model, the parallel corpora used was the combined isiZulu-English and isiXhosa-English corpora, with a language model trained only on the target language monolingual data. The number of additional sentences used from the other Nguni language was equivalent to the number of synthetic parallel corpora sentences generated via back-translation. For example, in the isiZulu backtranslation case 116618 synthetic parallel isiZulu-English sentences, as such in the multilingual system case 116618 isiXhosa-English sentences were added to the original isiZulu-English parallel corpora. This was done to allow for a comparison to be performed between the multilingual and back-translation systems, whilst avoiding a distortion caused by different sized corpora.

4 Test Design Method

4.1 Testing Data Sets

Evaluation data was sourced from the Autshumato project and consisted of 4 sets of candidate translations of an English to isiZulu set of sentences, as well as 4 sets of candidate translations of an English to isiXhosa set of sentences. This parallel corpora was different from any observed by the models during the training process, and consisted of 514 sentences.

Prior to being used for translations, all evaluation sets were normalised, and tokenised as outlined previously using byte-pair encoding tokenisation.

4.2 Translating Testing Data Sets

For each of the 3 model types, the English side corpora was given as input to the translation system. The translations were then formulated by the systems and written out into corresponding .txt files. This process was repeated for each set of parallel corpora for each Nguni language, resulting in four translation iterations for both isiZulu and isiXhosa.

4.3 Evaluating Translations

The BLEU (Bilingual Evaluation Understudy) metric was used as the evaluation metric. A python script using the NLTK Corpus BLEU module was used to perform this evaluation.

5 Results & Discussions

The results obtained for each of the three model types are outlined in the table below as the average BLEU score across the four evaluation translation sets.

Target Language	Baseline	Back-translation	Multilingual
IsiZulu	19.24	27.656	15.095
IsiXhosa	44.57	50.176	45.201

For the 3 translation systems created for isiZulu, the back-translation system had the best performance with a BLEU score of 27.656 out of a 100, followed by the baseline system with a score of 19.24 and finally the multilingual system with a score of 15.095. In this case the results were contrary to the original hypothesis presented, for the Zulu dataset, the multilingual system had the worst performance despite boasting 116618 more parallel sentences than the baseline system, with the back-translation system yielding the best result.

Similarly, of the three translation system created for isiXhosa, the back-translation system yielded the best performance with a BLEU score of 50.176, however in this case it was followed by the multilingual system with a BLEU score of 45.201 and finally the baseline system with a score of 44.57. In this case, whilst the hypothesis that the multilingual system would yield the best results was proven false, the two systems that made use of data augmentation techniques did yield better BLEU scores than the baseline system.

Additionally, on average the use of data augmentation techniques led to increased machine translation quality in the low-resource Nguni language setting, as can be seen by the back-translation system outperforming the baseline system in both the isiZulu and isiXhosa context, whilst multilingual translation also outperformed the baseline system in the isiXhosa setting. However, the multilingual systems poor performance in the isiZulu context may also indicate the need for some minimum threshold to be met with regards to the number of original parallel corpora required to allow for multilingual data to yield improvements in translation quality for a specific target language.

Finally, the isiXhosa translation systems had significantly better BLEU score across all three types of systems than the isiZulu translation systems, further indicating the importance of training dataset size to the performance of the SMT system.

6 Conclusions & Future Work

The following conclusions can thus be reached as a result of this project:

- On average, making use of data augmentation techniques to increase the available corpora in the low-resource Nguni language translation setting yields better machine translation quality than using the limited available original parallel corpora alone for training purposes.

- Regardless of corpora size, as shown in both the isiZulu and isiXhosa contexts the back-translations systems outperformed both the multilingual and baseline systems. As such it can be concluded that using synthetic parallel corpora, generated from existing parallel corpora, as additional data for Nguni language machine translation is a better data augmentation approach to use when compared to the use of multilingual parallel training data.
- Multilingual data augmentation displays inconsistencies in performance that may indicate that there exits some optimal mix of original parallel corpora and multilingual corpora in order to lead to improvements in translation quality.

The conclusions outlined above indicate that there is value in generating and collecting increasingly larger sets of monolingual data in Nguni languages that can later be used to generate synthetic corpora to be used in the machine translation training process. Additionally, they provide motivation for further studies to be conducted into the use of more complex data augmentation techniques in SMT for Nguni languages. As such, possible future work in this research area could seek to investigate whether there exists some threshold where synthetic corpora ceases to yield improvements to the Nguni machine translation systems, whilst also allowing for the development of systems dedicated to finding web-based monolingual Nguni language data. Additionally, further research could be conducted in the realm of multilingual data augmentation for Nguni languages with the use of larger data sets or using more than two target languages sets in the training process.

References

[1] Agić, Ž., and Vulić, I. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 3204–3210.

[2] Aharoni, R., Johnson, M., and Firat, O. Massively Multilingual Neural Machine Translation. *arXiv:1903.00089 [cs]* (July 2019). arXiv: 1903.00089.

[3] Akshai Ramesh, Venkatesh Balavadhani, Rejwanul Haque, and Andy Way. Investigating Low-resource Machine Translation for English-to-Tamil.

[4] Bertoldi, N., and Federico, M. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09* (Athens, Greece, 2009), Association for Computational Linguistics, p. 182.

[5] Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. A statistical approach to machine translation. *Comput. Linguistics 16* (1990), 79–85.

[6] Department of Arts and Culture, South Africa, and Centre for Text Technology, North-West University. Nchlt text corpora.

[7] Department of Arts and Culture, South Africa, and Centre for Text Technology, North-West University. Autshumato project data, 2007.

[8] Griesel, M., and McKellar, C. Syntactic Reordering as Pre-processing Step in Statistical Machine Translation of English to Sesotho sa Leboa and Afrikaans.

[9] Gu, J., Hassan, H., Devlin, J., and Li, V. O. K. Universal Neural Machine Translation for Extremely Low Resource Languages. *arXiv:1802.05368 [cs]* (Apr. 2018). arXiv: 1802.05368.

[10] Imankulova, A., Sato, T., and Komachi, M. Filtered Pseudo-parallel Corpus Improves Low-resource Neural Machine Translation. *ACM Transactions on Asian and Low-Resource Language Information Processing 19*, 2 (Mar. 2020), 1–16.

[11] Karakanta, A., Dehdari, J., and van Genabith, J. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation 32*, 1-2 (June 2018), 167–189.

[12] Keet, M., Mahlaza, Z., Heafield, K., Birch, A., and Pal, P. Medical machine translation, 2008.

[13] Koehn, P. *Statistical machine translation*. Cambridge University Press, Cambridge ; New York, 2010. OCLC: ocn316824008.

[14] Koehn, P. *Neural machine translation*, first edition ed. Cambridge University Press, New York, 2020.

[15] Koehn, P., and Knowles, R. Six Challenges for Neural Machine Translation. *arXiv:1706.03872 [cs]* (June 2017). arXiv: 1706.03872.

[16] Koehn, P., Och, F. J., and Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (2003), pp. 127–133.

[17] Laurette, M. Mburisano covid-19 multilingual corpus, 2020.

[18] Lopez, A. A survey of statistical machine translation.

[19] Martinus, L., and Abbott, J. Z. Benchmarking Neural Machine Translation for Southern African Languages. *arXiv:1906.10511 [cs, stat]* (June 2019). arXiv: 1906.10511.

[20] Mesham, S., Hayward, L., Shapiro, J., and Buys, J. Low-resource language modelling of south african languages.

[21] Poibeau, T. *Machine translation*. The MIT Press essential knowledge series. The MIT Press, Cambridge, Massachusetts, 2017.

[22] Resnik, P., and Smith, N. A. The Web as a Parallel Corpus. *Computational Linguistics 29*, 3 (Sept. 2003), 349–380.

[23] Rubino, R., Marie, B., Dabre, R., Fujita, A., Utiyama, M., and Sumita, E. Extremely low-resource neural machine translation for Asian languages. *Machine Translation 34*, 4 (Dec. 2020), 347–382.

[24] Sennrich, R., Haddow, B., and Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Berlin, Germany, 2016), Association for Computational Linguistics, pp. 86–96.

[25] Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., and Way, A. Human versus automatic quality evaluation of nmt and pbsmt. *Machine Translation 32* (2018), 217–235.

[26] Wang, R., Utiyama, M., Goto, I., Sumita, E., Zhao, H., and Lu, B.-L. Converting Continuous-Space Language Models into N -gram Language Models with Efficient Bilingual Pruning for Statistical Machine Translation. *ACM Transactions on Asian and Low-Resource Language Information Processing 15*, 3 (Mar. 2016), 1–26.