A Focus on Neural Machine Translation for African Languages

Abstract

African languages are numerous, complex and low-resourced. The datasets required for machine translation are difficult to discover, and existing research is hard to reproduce. Minimal attention has been given to machine translation for African languages so there is scant research regarding the problems that arise when using machine translation techniques. To begin addressing these problems, we trained models to translate English to five of the official South African languages (Afrikaans, isiZulu, Northern Sotho, Setswana, Xitsonga), making use of modern neural machine translation techniques. The results obtained show the promise of using neural machine translation techniques for African languages. By providing reproducible publicly-available data, code and results, this research aims to provide a starting point for other researchers in African machine translation to compare to and build upon.

Introduction

Africa has over 2000 languages across the continent BIBREF0 . South Africa itself has 11 official languages. Unlike many major Western languages, the multitude of African languages are very low-resourced and the few resources that exist are often scattered and difficult to obtain.

Machine translation of African languages would not only enable the preservation of such languages, but also empower African citizens to contribute to and learn from global scientific, social and educational conversations, which are currently predominantly English-based BIBREF1 . Tools, such as Google Translate BIBREF2 , support a subset of the official South African languages, namely English, Afrikaans, isiZulu, isiXhosa and Southern Sotho, but do not translate the remaining six official languages.

Unfortunately, in addition to being low-resourced, progress in machine translation of African languages has suffered a number of problems. This paper discusses the problems and reviews existing machine translation research for African languages which demonstrate those problems. To try to solve the highlighted problems, we train models to perform machine translation of English to Afrikaans, isiZulu, Northern Sotho (N. Sotho), Setswana and Xitsonga, using state-of-the-art neural machine translation (NMT) architectures, namely, the Convolutional Sequence-to-Sequence (ConvS2S) and Transformer architectures.

Section SECREF2 describes the problems facing machine translation for African languages, while the target languages are described in Section SECREF3 . Related work is presented in Section SECREF4 , and the methodology for training machine translation models is discussed in Section SECREF5 . Section SECREF6 presents quantitative and qualitative results.

Problems

The difficulties hindering the progress of machine translation of African languages are discussed below.

Low availability of resources for African languages hinders the ability for researchers to do machine translation. Institutes such as the South African Centre for Digital Language Resources (SADiLaR) are attempting to change that by providing an open platform for technologies and resources for South African languages BIBREF7 . This, however, only addresses the 11 official languages of South Africa and not the greater problems within Africa.

Discoverability: The resources for African languages that do exist are hard to find. Often one needs to be associated with a specific academic institution in a specific country to gain access to the language data available for that country. This reduces the ability of countries and institutions to combine their knowledge

and datasets to achieve better performance and innovations. Often the existing research itself is hard to discover since they are often published in smaller African conferences or journals, which are not electronically available nor indexed by research tools such as Google Scholar.

Reproducibility: The data and code of existing research are rarely shared, which means researchers cannot reproduce the results properly. Examples of papers that do not publicly provide their data and code are described in Section SECREF4 .

Focus: According to BIBREF8 , African society does not see hope for indigenous languages to be accepted as a more primary mode for communication. As a result, there are few efforts to fund and focus on translation of these languages, despite their potential impact.

Lack of benchmarks: Due to the low discoverability and the lack of research in the field, there are no publicly available benchmarks or leader boards to new compare machine translation techniques to.

This paper aims to address some of the above problems as follows: We trained models to translate English to Afrikaans, isiZulu, N. Sotho, Setswana and Xitsonga, using modern NMT techniques. We have published the code, datasets and results for the above experiments on GitHub, and in doing so promote reproducibility, ensure discoverability and create a baseline leader board for the five languages, to begin to address the lack of benchmarks.

Languages

We provide a brief description of the Southern African languages addressed in this paper, since many readers may not be familiar with them. The isiZulu, N. Sotho, Setswana, and Xitsonga languages belong to the Southern Bantu group of African languages BIBREF9 . The Bantu languages are agglutinative and

all exhibit a rich noun class system, subject-verb-object word order, and tone BIBREF10 . N. Sotho and Setswana are closely related and are highly mutually-intelligible. Xitsonga is a language of the Vatsonga people, originating in Mozambique BIBREF11 . The language of isiZulu is the second most spoken language in Southern Africa, belongs to the Nguni language family, and is known for its morphological complexity BIBREF12 , BIBREF13 . Afrikaans is an analytic West-Germanic language, that descended from Dutch settlers BIBREF14 .

Related Work

This section details published research for machine translation for the South African languages. The existing research is technically incomparable to results published in this paper, because their datasets (in particular their test sets) are not published. Table TABREF1 shows the BLEU scores provided by the existing work.

Google Translate BIBREF2 , as of February 2019, provides translations for English, Afrikaans, isiZulu, isiXhosa and Southern Sotho, six of the official South African languages. Google Translate was tested with the Afrikaans and isiZulu test sets used in this paper to determine its performance. However, due to the uncertainty regarding how Google Translate was trained, and which data it was trained on, there is a possibility that the system was trained on the test set used in this study as this test set was created from publicly available governmental data. For this reason, we determined this system is not comparable to this paper's models for isiZulu and Afrikaans.

BIBREF3 trained Transformer models for English to Setswana on the parallel Autshumato dataset BIBREF15 . Data was not cleaned nor was any additional data used. This is the only study reviewed that released datasets and code. BIBREF4 performed statistical phrase-based translation for English to Setswana translation. This research used linguistically-motivated pre- and post-processing of the corpus

in order to improve the translations. The system was trained on the Autshumato dataset and also used an additional monolingual dataset.

BIBREF5 used statistical machine translation for English to Xitsonga translation. The models were trained on the Autshumato data, as well as a large monolingual corpus. A factored machine translation system was used, making use of a combination of lemmas and part of speech tags.

BIBREF6 used unsupervised word segmentation with phrase-based statistical machine translation models. These models translate from English to Afrikaans, N. Sotho, Xitsonga and isiZulu. The parallel corpora were created by crawling online sources and official government data and aligning these sentences using the HunAlign software package. Large monolingual datasets were also used.

BIBREF16 performed word translation for English to isiZulu. The translation system was trained on a combination of Autshumato, Bible, and data obtained from the South African Constitution. All of the isiZulu text was syllabified prior to the training of the word translation system.

It is evident that there is exceptionally little research available using machine translation techniques for Southern African languages. Only one of the mentioned studies provide code and datasets for their results. As a result, the BLEU scores obtained in this paper are technically incomparable to those obtained in past papers.

Methodology

The following section describes the methodology used to train the machine translation models for each language. Section SECREF4 describes the datasets used for training and their preparation, while the algorithms used are described in Section SECREF8 .

Data

The publicly-available Autshumato parallel corpora are aligned corpora of South African governmental data which were created for use in machine translation systems BIBREF15 . The datasets are available for download at the South African Centre for Digital Language Resources website. The datasets were created as part of the Autshumato project which aims to provide access to data to aid in the development of open-source translation systems in South Africa.

The Autshumato project provides parallel corpora for English to Afrikaans, isiZulu, N. Sotho, Setswana, and Xitsonga. These parallel corpora were aligned on the sentence level through a combination of automatic and manual alignment techniques.

The official Autshumato datasets contain many duplicates, therefore to avoid data leakage between training, development and test sets, all duplicate sentences were removed. These clean datasets were then split into 70% for training, 30% for validation, and 3000 parallel sentences set aside for testing. Summary statistics for each dataset are shown in Table TABREF2 , highlighting how small each dataset is.

Even though the datasets were cleaned for duplicate sentences, further issues exist within the datasets which negatively affects models trained with this data. In particular, the isiZulu dataset is of low quality. Examples of issues found in the isiZulu dataset are explained in Table TABREF3 . The source and target sentences are provided from the dataset, the back translation from the target to the source sentence is given, and the issue pertaining to the translation is explained.

Algorithms

We trained translation models for two established NMT architectures for each language, namely, ConvS2S and Transformer. As the purpose of this work is to provide a baseline benchmark, we have not performed significant hyperparameter optimization, and have left that as future work.

The Fairseq(-py) toolkit was used to model the ConvS2S model BIBREF17 . Fairseq's named architecture “fconv” was used, with the default hyperparameters recommended by Fairseq documentation as follows: The learning rate was set to 0.25, a dropout of 0.2, and the maximum tokens for each mini-batch was set to 4000. The dataset was preprocessed using Fairseq's preprocess script to build the vocabularies and to binarize the dataset. To decode the test data, beam search was used, with a beam width of 5. For each language, a model was trained using traditional white-space tokenisation, as well as byte-pair encoding tokenisation (BPE). To appropriately select the number of tokens for BPE, for each target language, we performed an ablation study (described in Section SECREF25).

The Tensor2Tensor implementation of Transformer was used BIBREF18 . The models were trained on a Google TPU, using Tensor2Tensor's recommended parameters for training, namely, a batch size of 2048, an Adafactor optimizer with learning rate warm-up of 10K steps, and a max sequence length of 64. The model was trained for 125K steps. Each dataset was encoded using the Tensor2Tensor data generation algorithm which invertibly encodes a native string as a sequence of subtokens, using WordPiece, an algorithm similar to BPE BIBREF19 . Beam search was used to decode the test data, with a beam width of 4.

Results

Section SECREF9 describes the quantitative performance of the models by comparing BLEU scores, while a qualitative analysis is performed in Section SECREF10 by analysing translated sentences as well as attention maps. Section SECREF25 provides the results for an ablation study done regarding the

effects of BPE.

Quantitative Results

The BLEU scores for each target language for both the ConvS2S and the Transformer models are reported in Table TABREF7 . For the ConvS2S model, we provide results for sentences tokenised by white spaces (Word), and when tokenised using the optimal number of BPE tokens (Best BPE), as determined in Section SECREF25 . The Transformer model uses the same number of WordPiece tokens as the number of BPE tokens which was deemed optimal during the BPE ablation study done on the ConvS2S model.

In general, the Transformer model outperformed the ConvS2S model for all of the languages, sometimes achieving 10 BLEU points or more over the ConvS2S models. The results also show that the translations using BPE tokenisation outperformed translations using standard word-based tokenisation. The relative performance of Transformer to ConvS2S models agrees with what has been seen in existing NMT literature BIBREF20 . This is also the case when using BPE tokenisation as compared to standard word-based tokenisation techniques BIBREF21 .

Overall, we notice that the performance of the NMT techniques on a specific target language is related to both the number of parallel sentences and the morphological typology of the language. In particular, isiZulu, N. Sotho, Setswana, and Xitsonga languages are all agglutinative languages, making them harder to translate, especially with very little data BIBREF22 . Afrikaans is not agglutinative, thus despite having less than half the number of parallel sentences as Xitsonga and Setswana, the Transformer model still achieves reasonable performance. Xitsonga and Setswana are both agglutinative, but have significantly more data, so their models achieve much higher performance than N. Sotho or isiZulu.

The translation models for isiZulu achieved the worst performance when compared to the others, with the maximum BLEU score of 3.33. We attribute the bad performance to the morphological complexity of the language (as discussed in Section SECREF3), the very small size of the dataset as well as the poor quality of the data (as discussed in Section SECREF4).

Qualitative Results

We examine randomly sampled sentences from the test set for each language and translate them using the trained models. In order for readers to understand the accuracy of the translations, we provide back-translations of the generated translation to English. These back-translations were performed by a speaker of the specific target language. More examples of the translations are provided in the Appendix. Additionally, attention visualizations are provided for particular translations. The attention visualizations showed how the Transformer multi-head attention captured certain syntactic rules of the target languages.

In Table TABREF20 , ConvS2S did not perform the translation successfully. Despite the content being related to the topic of the original sentence, the semantics did not carry. On the other hand, Transformer achieved an accurate translation. Interestingly, the target sentence used an abbreviation, however, both translations did not. This is an example of how lazy target translations in the original dataset would negatively affect the BLEU score, and implore further improvement to the datasets. We plot an attention map to demonstrate the success of Transformer to learn the English-to-Afrikaans sentence structure in Figure FIGREF12 .

Despite the bad performance of the English-to-isiZulu models, we wanted to understand how they were performing. The translated sentences, given in Table TABREF21 , do not make sense, but all of the words are valid isiZulu words. Interestingly, the ConvS2S translation uses English words in the

translation, perhaps due to English data occurring in the isiZulu dataset. The ConvS2S however correctly prefixed the English phrase with the correct prefix “i-". The Transformer translation includes invalid acronyms and mentions “disease" which is not in the source sentence.

If we examine Table TABREF22 , the ConvS2S model struggled to translate the sentence and had many repeating phrases. Given that the sentence provided is a difficult one to translate, this is not surprising. The Transformer model translated the sentence well, except included the word “boithabišo”, which in this context can be translated to “fun” - a concept that was not present in the original sentence.

Table TABREF23 shows that the ConvS2S model translated the sentence very successfully. The word “khumo” directly means “wealth” or “riches”. A better synonym would be “letseno”, meaning income or “letlotlo” which means monetary assets. The Transformer model only had a single misused word (translated “shortage” into “necessity”), but otherwise translated successfully. The attention map visualization in Figure FIGREF18 suggests that the attention mechanism has learnt that the sentence structure of Setswana is the same as English.

An examination of Table TABREF24 shows that both models perform well translating the given sentence. However, the ConvS2S model had a slight semantic failure where the cause of the economic growth was attributed to unemployment, rather than vice versa.

Ablation Study over the Number of Tokens for Byte-pair Encoding

BPE BIBREF21 and its variants, such as SentencePiece BIBREF19 , aid translation of rare words in NMT systems. However, the choice of the number of tokens to generate for any particular language is not made obvious by literature. Popular choices for the number of tokens are between 30,000 and 40,000: BIBREF20 use 37,000 for WMT 2014 English-to-German translation task and 32,000 tokens for the WMT

2014 English-to-French translation task. BIBREF23 used 32,000 SentencePiece tokens across all source and target data. Unfortunately, no motivation for the choice for the number of tokens used when creating sub-words has been provided.

Initial experimentation suggested that the choice of the number of tokens used when running BPE tokenisation, affected the model's final performance significantly. In order to obtain the best results for the given datasets and models, we performed an ablation study, using subword-nmt BIBREF21 , over the number of tokens required by BPE, for each language, on the ConvS2S model. The results of the ablation study are shown in Figure FIGREF26 .

As can be seen in Figure FIGREF26 , the models for languages with the smallest datasets (namely isiZulu and N. Sotho) achieve higher BLEU scores when the number of BPE tokens is smaller, and decrease as the number of BPE tokens increases. In contrast, the performance of the models for languages with larger datasets (namely Setswana, Xitsonga, and Afrikaans) improves as the number of BPE tokens increases. There is a decrease in performance at 20 000 BPE tokens for Setswana and Afrikaans, which the authors cannot yet explain and require further investigation. The optimal number of BPE tokens were used for each language, as indicated in Table TABREF7 .

Future Work

Future work involves improving the current datasets, specifically the isiZulu dataset, and thus improving the performance of the current machine translation models.

As this paper only provides translation models for English to five of the South African languages and Google Translate provides translation for an additional two languages, further work needs to be done to provide translation for all 11 official languages. This would require performing data collection and

incorporating unsupervised BIBREF24 , BIBREF25 , meta-learning BIBREF26 , or zero-shot techniques BIBREF23 .

Conclusion

African languages are numerous and low-resourced. Existing datasets and research for machine translation are difficult to discover, and the research hard to reproduce. Additionally, very little attention has been given to the African languages so no benchmarks or leader boards exist, and few attempts at using popular NMT techniques exist for translating African languages.

This paper reviewed existing research in machine translation for South African languages and highlighted their problems of discoverability and reproducibility. In order to begin addressing these problems, we trained models to translate English to five South African languages, using modern NMT techniques, namely ConvS2S and Transformer. The results were promising for the languages that have more higher quality data (Xitsonga, Setswana, Afrikaans), while there is still extensive work to be done for isiZulu and N. Sotho which have exceptionally little data and the data is of worse quality. Additionally, an ablation study over the number of BPE tokens was performed for each language. Given that all data and code for the experiments are published on GitHub, these benchmarks provide a starting point for other researchers to find, compare and build upon.

The source code and the data used are available at https://github.com/LauraMartinus/ukuxhumana.

Acknowledgements

The authors would like to thank Reinhard Cromhout, Guy Bosa, Mbongiseni Ncube, Seale Rapolai, and Vongani Maluleke for assisting us with the back-translations, and Jason Webster for Google Translate

API assistance. Research supported with Cloud TPUs from Google's TensorFlow Research Cloud (TFRC).

Appendix

Additional translation results from ConvS2S and Transformer are given in Table TABREF27 along with their back-translations for Afrikaans, N. Sotho, Setswana, and Xitsonga. We include these additional sentences as we feel that the single sentence provided per language in Section SECREF10 , is not enough demonstrate the capabilities of the models. Given the scarcity of research in this field, researchers might find the additional sentences insightful into understanding the real-world capabilities and potential, even if BLEU scores are low.