Machine Translation for an English-to-isiZulu Parallel Corpus

Olatomiwa Akinlaja
Computer Sci. & Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
2534648

Kudzai Sibanda
Computer Sci. & Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
935227

Abstract

We are living in a digital age where connecting with other people is an important part of staying alive. We interact with people from different places who have different cultures and languages and sometimes language can be a barrier. The introduction of translations helped to bridge the gap between speakers of different languages to achieve a common goal. Translations have been made for many languages in the world and they continue to devise new methods of collecting data and forming translations for more languages. The application of machine translation and NLP techniques for African languages have become valuable for the progress of African NLP and has been very useful in the development of translation models and dictionaries to help with translation. In this project we attempt to perform translations of English to Zulu parallel corpus using three different models to see how well they performed and the translation they could produce. We implemented a custom seq2seq model, JoeyNMT as well as a Huggingface model.

1 Problem Definition/Introduction

Natural language processing has become an important tool when it comes to globalization and dealing with languages. A lot of tools have been developed for language translation to aid in interactions between people from different countries as well as translation of texts so people can read books/articles/publications in their own languages and not just limited to English. It is therefore important for the translation tools used to perform at a high level so as to prevent errors or miscommunication. There are various NLP techniques that have been developed for language translation and they all perform differently depending on the language they are translating. Language translation is a costly process and this led to the development of automatic machine translation systems which still have not reached the level of human translators. Tokenization is an essential part of a machine's process of learning a language. It is the breaking down of words into smaller units (tokens) to allow the computer to learn and build a vocabulary that will be instrumental in the translation process. With the various types of translation techniques available, which is best for English to Zulu translation? What machine translation techniques are suitable for English to Zulu translation?

2 Background

Translation of text into African languages became an important task in recent years and it has also proven to be a very difficult task due to the complex nature of the languages. Unlike English, African languages present a morphological complexity in which a full expression/sentence can be written as a single word. Zulu, one of the official languages in South Africa, is the focus of this study and like other African languages it has been worked on by some NLP enthusiasts and they published various projects to present their findings.

(Wolff and Kotzé, 2014) experimented with Zulu to English translation using syllables as tokens with BLEU measures for evaluation. They used a statistical approach to their training of a machine translator, however this method requires a large amount of data for training so the model can learn as much as it can before having to be tested. They used a Zulu Bible and a King James Bible for their text. Their approach seemed to have promising results and in some cases provided better translations than word-based translators .

(Pretorius and Bosch, 2002) explore the use of tools for Zulu language processing (XML and finite-state tools) and the important role that regular expressions play in the translation process. They provide insight into the main components of the processing of Zulu which are lexicon and morphological analysis. They recommend the use of automated lexicon analyzers for language translation. They also worked on a way to enhance ZulMorph (morphological analyser for Zulu) by using English as a pivot language for the Zulu semantics. They present a Zulu lexical- knowledge base (Bosch and Pretorius, 2017) which is well resourced to deal with the issue of giving meaning to the verbs being analysed by ZulMorph. They also worked on a corpus-based method which uses NLP tools to create new word and verb roots and add more lexicalisations to the Zulu lexical-knowledge base.

Despite the growing work in the translation world there is still a concern on how low-resourced African languages are and how that poses a challenge when it comes to accurate translation. Many NLP enthusiasts have begun the work of putting together dictionaries that can be used to facilitate the translation process. (Bosch and Faaß, 2014) described the very first type of an e-dictionary that translates possessive constructs from English into Zulu. Their approach used a combination of e-lexicography and a web dictionary and this allowed the creation of a dictionary. They built a prototype dictionary ,albeit with a few struggles with Zulu corpus, from their research which is able to translate possessives from English to Zulu and this is major progress in the Zulu language translation process.

(Munteanu and Marcu, 2005) worked on a way to improve the performance of Machine Translation models by finding parallel sentences in non-parallel datasets with the help of a maximum entropy classifier. The classifier is to predict whether a pair of sentences are translations of each other. This method was introduced to aid with the development of low resource languages by increasing the corpora sizes. They further claim that the greatest challenge of machine translation is the lack of parallel corpora for most languages and the goal is to create ways of automating the acquisition of the parallel corpora.

3 Methodology

The data used for training is a parallel corpus within a Comma Separated Value (csv) file which has 3 columns (English sentences, Zulu sentences and the source of the sentences).

The first model is a machine translation that uses Byte Pair Encoding (BPE) tokenization and BLEU scores for evaluation. Byte pair encoding is a type of tokenization which was introduced in 1994 (Provilkov et al., 2019) as a method of compressing data that replaced the pair most frequent in a sequence with an unused byte and now it is used for splitting rare and unknown words and keeps the frequent ones. The modern version of BPE is likened to an adaptation of a word segmentation algorithm. Its algorithm is divided into two phases which are training and inferring.

The training phase is where the algorithm segments each word in the sentence as a sequence of characters and this allows for the development of a merge table and a vocabulary for the tokens (Voita, 2021). The most frequently appearing pairs of words are merged into the merge table and the token is added to the vocabulary and this process is carried out until all the words in the corpus have been counted and checked for frequency. The inference phase is where the algorithm uses the merge table created in the training phase to segment new text (Voita, 2021). This phase is mainly about the algorithm finding the highest merge in the table after and applying it until there are no more possible merges left. The merge table arranges the merges in order of their frequency with the most frequent appearing at the top.

BLEU score (BiLingual Evaluation Understudy) is a tool for evaluating text in machine translation model results. The score is presented as a value between zero and one and it is a measure of how similar a model's translation prediction is in relation to human translations, with a score of zero being a low quality translation with no overlap between the machine translation and the reference translation and one is the high quality translation that has a perfect match between the machine translation and the reference translation (Cloud, 2021). BLEU only performs well when used to evaluate corpuses and it can not be used for individual sentences. It also does not capture meaning or grammar but simply focuses on the corpus as a whole. Tokenization and normalization have to be done before BLEU scores are calculated.

The first model uses BPE on a custom parallel corpus using the JoeyNMT framework.The base code is freely available from Masakhane who do machine translation for African languages and promote research into NLP for African translations. The results are presented in the Results section of this paper.

The second model is a Hugging Face pre-trained model. The goal was to create a translation model with multilingual encoder/decoder that is able to function without relying on English data for training.

Encoder - Decoder framework is a well known modelling tool used when working on sequence-to-sequence tasks. It is divided into the encoding stage and the decoding stage (Voita, 2021). The encoding stage is where the source of the data is read and the encoder produces a source representation and passes it onto the decoder. The decoder takes the source representation and generates a target sequence.

The steps involved; Importing the pipeline function and initialising the pipeline function with the required parameter, which are specifying the task and pre-trained model. The last step involves performing language translation (English-isiZulu).

In order to perform machine translation, we have to provide a token ID, and tokenized attribute in order to specify the language we would like to translate to. We modify the translation even further by calculating the BLEU score for each translation in order to get a general perspective of how accurate the translation is.

4 Results

4.1 Attempted starter notebook with custom parallel corpus, JoeyNMT

Performing machine translation using the English to isiZulu parallel corpus, incorporates BPE tokenization. An evaluation of the model via the BLEU score. The model performed poorly and this is because of the small size of the training data. A small sample of the results is presented below:

4.1.1 Source/Reference/Hypothesis

- Source: Conor Garland added his team-leading 11th goal in the third period, and Clayton Keller notched two assists for the Coyotes, who leapfrogged idle Edmonton into first place in the Pacific Division. Arizona has won the first two contests of its four-game road trip and improved to 10-3-3 away from home this season.

- Reference: UConor Garland ufake igoli le-11 leqembu lakhe ebelihamba phambili ehlandleni lesithathu, bese uClayton Keller wazisa kabili kumaCoyote, ogxumise i-Edmonton yaba sendaweni yokuqala Ophikweni lakuPacific. I-Arizona inqobe imincintiswano emibili yokuqala ohambweni lwayo lwemidlalo emine futhi yathuthukela ku-10-3-3 engakho ekhaya kulesi sikhathi sokudlala.

- Hypothesis: UUUukuthi ukuthi ukuthi ukuthi ukuthi

- BLEU Score: 0.02

5 Attempted Hugging Face pre trained model:

Specified text2text generation, The pre-trained model is downloaded from Facebook, called m2m100. A multilingual encoder/decoder model, a sequence to sequence model trained for many to many multilingual translations. The model can translate to many different languages without relying on English data.

Steps involved: Import the pipeline function and initialising the pipeline function with the required parameter, which are specifying the task and pre-trained model. The last step involves performing language translation (eng-isizulu).

In order to perform machine translation, we have to provide a token ID, and tokenized attribute in order to specify the language we would like to translate to. We modify the translation even further by calculating the BLEU score for each translation in order to get a general perspective of how accurate the translation is.

Huggingface performed better than the previous two models:

5.0.1 Source/Reference/Hypothesis

- Source: Peter Van Sant: And it means what?
- Reference: Peter Van Sant: Bese kusho ukuthini?
- Hypothesis: I-Peter Van Sant: Futhi lokhu kuyinto?
- BLEU Score: 0.508133

6 Attempted seq2seq model

We implemented a seq2seq technique for language translation (Voita, 2021). The English and isi-Zulu corpuses were combined in order to feed it into the model. We decided to test out the performance of the seq2seq model without the use of the attention optimization. The model did not perform well and the predictions it made were not accurate and the metrics were low.

6.0.1 Source/Reference/Hypothesis

- Source: According to a former Fox executive, Ailes then blew up at Bill Shine, who had authorized Hannity's trip.
- Reference: Ngokwalowo owayeyisikhulu esiphezulu se-Fox, u-Ailes waxabana noBill Shine, owayegunyaze uhambo lukaHannity.
- Hypothesis: kodwa ukuthi ukuthi ukuthi ukuthi oneminyaka esiphezulu futhi yokuthi abasebenzi abasebenzi abasebenzi izidakamizwa esiphezulu james asibona Bleu
- BLEU Score: 0.06213502070319737

7 Impact Statement

7.1 How could your research affect ML applications?

The development of a machine translation model for English to Zulu sheds more light on the need for ML applications to be developed for the betterment of understanding African languages. These languages have been overlooked for decades while NLP has been used to translate languages such as French, Portuguese and other European languages. The model helps enhance our knowledge of language translation using NLP methods. It brings about a need for the creation of more of these types of applications. This model adds to the body of knowledge and the list of NLP translation models for English-Zulu. This research has begun the use of sequence to sequence machine translation and this is a new addition to the Zulu translation models and it provides a chance for other researchers to enhance and make better predictions and translations. The research does not have any ethical concerns. The data was publically available for analysis and it does not contain any sensitive or personal information. The results do not endanger or bring any individuals any harm. The data is not biased in any way.The data is simply conversational sentences. The model is complex to understand but it can be explained in terms of its functioning. The performance of the model is represented in the form of BLEU scores and they correctly represent the model. We used BLEU scores because that is the most appropriate tool to measure the performance of a machine translation model. The building of the model does not require a complex system and it can be done with any computer that has access to a GPU. There is no environmental impact to creating the model. The model can be adapted to the translation of other languages with a few tweaks to the code to suit the morphology of the language in question.

7.2 What are the societal implications of these applications?

An English-Zulu translation model can be used in various ways outside of the lab. The easiest way to implement it would be to embed it in a translation machine for tourists that are visiting areas where Zulu is spoken e.g. Kwa-Zulu Natal. It would make it easier for them to communicate with locals and get a better experience. Another possible use would be using it as an English-Zulu dictionary application for translations and it can be used by anyone who can install it on their devices. The malicious uses of the model could be using the translation to take advantage of Zulu speakers . It provides a way for criminals to be able to communicate with Zulu speakers and take advantage of them.

The products created from the model can be used in various industries eg. tourism, education and even retail. It has mostly beneficial and positive impacts but it can also have negative impacts if used with malicious intentions. In the tourism industry it can be used to enhance experiences by allowing them to interact with locals and not just tour guides. In the educational sector it can be used as a dictionary for students that study in isiZulu. In the legal sector it can be used to communicate with witnesses and even be used during prosecution and avoid having to use human translators.

7.3 What research or other initiatives could improve societal outcomes?

ML research suggestions - More work needs to be done in terms of African language translations. Africa has a lot of languages with diverse attributes and they need to be studied and translated too. Language translation needs to be used on more than just the 'big languages' but also the lesser known ones. Development of translation models that can work with low resourced languages is needed and this will help improve the societal impacts mentioned in the previous section. Natural Language Processing, along with Machine Learning should be invested into and should be used to create useful automations and other useful tools that can make our lives easier.

7.4 Conclusion

In the beginning of the project we set out to create machine translation models and to evaluate their translations using BLEU scores. The first model was a Sequence-to-sequence translation model without Attention. It was an encoder-decoder model with LSTM neural machine translation using keras. It did not perform as well as we had thought it would but it showed a bit of promise with the translations.

The next model used JoeyNMT using a parallel corpus and BPE tokenization. Using Masakhane baseline code and encoder-decoder model we trained the English to Zulu corpus and used BLEU scores for performance measure. A HuggingFace model with the use of Wolf transformers was developed. It posed a challenge because it takes quite a long time to train and test and without a large enough corpus for training it is not able to perform as well as it can.

In conclusion, we used a few different translation models and because of the low resource nature of the data we were not able to get the high quality translations we had hoped. Machine translation is a complex process and there are many considerations that must be focused on in order to create a fully functional translator. It is possible to create a translation model for English to Zulu. The use of pre-trained models makes it easier to work on the translations. We hope that in future we can be able to use these models on datasets with larger corpuses and higher quality Zulu translations.

References

Sonja E. Bosch and Gertrud Faaß. 2014. Towards an Integrated e-Dictionary Application: The Case of an English to Zulu Dictionary of Possessives. In *Proceedings of the 16th Euralex Conference. Proceedings of the Sixteenth EURALEX International Congress, EURALEX 2014, Bolzano/Bozen, Italy, July 15–19*, pages 739–747.

Sonja E. Bosch and Laurette Pretorius. 2017. A computational approach to Zulu verb morphology within the context of lexical semantics. *Lexikos*, 27:152–182.

Google Cloud. 2021. Understanding BLEU scores.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Laurette Pretorius and Sonja Bosch. 2002. Regular expressions: enabling the development of computational aids for Zulu natural language processing. In *ACM International Conference Proceeding Series*, volume 30, pages 254–254.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*.

Elena Voita. 2021. Sequence to Sequence (seq2seq) and Attention.

Friedel Wolff and Gideon Kotzé. 2014. Experiments with syllable-based Zulu-English machine translation. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*, pages 217–222.