Towards Neural Machine Translation for African Languages

Jade Z. Abbott
Retro Rabbit
ja@retrorabbit.co.za

Laura Martinus
Human Language Technologies, CSIR
lmartinus@csir.co.za

Abstract

Given that South African education is in crisis, strategies for improvement and sustainability of high-quality, up-to-date education must be explored. In the migration of education online, inclusion of machine translation for low-resourced local languages becomes necessary. This paper aims to spur the use of current neural machine translation (NMT) techniques for low-resourced local languages. The paper demonstrates state-of-the-art performance on English-to-Setswana translation using the Autshumato dataset. The use of the Transformer architecture beat previous techniques by 5.33 BLEU points. This demonstrates the promise of using current NMT techniques for African languages.

1 Introduction

Given that South Africa's education system is in crisis [1], strategies for sustainability within education must be explored. One suggested strategy to improve youth education would be to augment the learning process with online content [2].

The internet comprises of 53.5% English content, while the other 10 official South African languages comprise of less than 0.1% of the languages spoken on the internet [3]. According to the 2011 South African census, only 9.8% of South Africans speak English as a primary language [4]. Similar statistics exist for many other African countries [5–7].

For the South African youth to benefit from the massive amount of educational online content, the translation of online resources into the many low-resourced African languages is sorely needed. Unfortunately, machine translation of low-resourced languages has proven difficult with both conventional statistical machine translation (SMT) and the more recent neural machine translation (NMT) methods [8–11].

The convolutional sequence-to-sequence (ConvS2S) architecture improved translation results on multiple languages, including low-resourced languages [12]. Additionally, by pre-training the Transformer architecture on many languages and then specialising to a single language, Gu *et al.* were able to improve performance on low-resourced languages [13].

This paper aims to serve as the initial work towards using modern neural machine translation (NMT) techniques to improve machine translation on African languages, and invigorate future research into using such techniques. NMT techniques are often overlooked in favour of conventional phrase-based translation systems. This is due to vanilla NMT's reputation for comparatively under-performing on low-resource languages [13]. This paper shows the performance of training convolutional sequence-to-sequence learning [12] and the Transformer architecture on the Autshumato English-Setswana Parallel Corpora [14].

Presented at NIPS 2018 Workshop on Machine Learning for the Developing World

2 Related Work

Kato *et al.* used statistical phrase-based translation, based on Moses, in order to perform English-to-Setswana translation [10]. They achieve a BLEU score of 32.71 on a dataset that is not publically-available and so was excluded from the comparison. Wilken *et al.* used a similar technique as [10], but focused on linguistically-motivated pre- and post-processing of the corpus in order to improve translation with phrase-based techniques [11]. Wilken *et al.* was trained on the same Autshumato dataset used in this paper, and also used an additional monolongual dataset for language modelling.

3 Methodology

Section 3.1 describes the parallel dataset used, while the selected models and their hyperparameters for training are described in Section 3.2.

3.1 Dataset

The publically-available Autshumato English-Setswana Parallel Corpora is an aligned corpus of South African governmental data which was created for the use in machine translation systems. The dataset consists of three smaller parallel corpora obtained from different sources which were combined to form a single corpus. The combined corpus was sharded into 111 300 sentences as training data, 44 700 as validation data, and 3 000 sentences set aside as test data for evaluation. This dataset is available for download at the South African Centre for Digital Language Resources website.[1]

3.2 Models

Limited work has been done using NMT techniques for African languages. In fact, as far as the authors can tell, this is the first work using modern NMT techniques for translation for South African official languages. We thus selected two recent NMT architectures, convolutional sequence-to-sequence and Transformer, to compare to existing research. The existing research uses phrase-based SMT for English-to-Setswana translation.

The Fairseq(-py) tookit was used to model the convolutional sequence to sequence model [12]. The model used was one of Fairseq's named architectures "fconv". 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.

The Tensor2Tensor implementation of Transformer was used [15]. The model was trained for 125K steps . The learning rate was set to 0.4, with a batch size of 1024, and a learning rate warm-up of 45000 steps. The dataset was encoded using the Tensor2Tensor data generation algorithm which invertibly encodes a native string as a sequence of subtokens [15]. Beam search was used to decode the test data, with a beam width of 4.

Training took less than 12 hours for both algorithms on a NVIDIA K80 GPU.

4 Results

Section 4.1 describes the quantitative performance of the models by comparing BLEU scores, while a brief qualitative analysis is performed in in Section 4.2.

4.1 Quantitative

According to the BLEU scores reported in Table 1, ConvS2S achieved a BLEU score of 27.77, which is 1.02 BLEU points below the performance of the phrase-based English-Setswana system from [11]. The Transformer model significantly outperformed the phrase-based English-Setswana system, by 5.33 BLEU points. Despite the fact neither ConvS2S nor Transformer had information

[1] Available online at: https://rma.nwu.ac.za/index.php/resource-catalogue/autshumato-english-setswana-multi-bilingual-corpus.html

Table 1: BLEU scores of the trained models on the Autshumato dataset.

Model	BLEU
From [11]	28.8
ConvS2S	27.77
Transformer (uncased)	33.53
Transformer (cased)	33.12

Table 2: For each source sentence we show the reference translation, and the translations by the various models. We also show the translation of the results back to English, performed by a Setswana speaker.

Source	Take a shower rather than a bath as showers use less water and electricity.
Reference	Dirisa šawara boemong jwa go tsena mo bateng ka gonne dišawara di dirisa eneji e nnye.
ConvS2S	Tsaya tsia go na le kgonagalo ya gore go na le metsi a a ka nnang dirisang metsi le motlakase.
Setswana-to-Eng Speaker ConvS2S	Take caution of the possibility that there is water that can remain use water and electricity.
Transformer	Tsamaya matsidinyana ka bonako go na le dibata ka di dirisa metsi a a kwa tlase le motlakase.
Setswana-to-Eng Speaker Transformer	Go really fast instead of baths because they use minimal water and electricity.
Source	This is to protect the abuse of children and young workers.
Reference	Ntlha e e botlhokwa go sireletsa tiriso ya bana le baŝwa.
ConvS2S	Se ke go sireletsa tshotlako ya bana le badiri ba bannye.
Setswana-to-Eng Speaker ConvS2S	This is to protect the abuse of children and young workers.
Transformer	Se ke go sireletsa tshotlako ya bana le baŝwa.
Setswana-to-Eng Speaker Transformer	This is to protect the abuse of children and adolescents.
Source	An example is a rural job creation project that aimed to deal with the fact that 40% of people in a community are unemployed.
Reference	Sekao ke porojeke ya go tlholwa ga ditiro kwa magaeng go samagana le ntlha ya gore 40% ya batho mo setšhabeng ga ba dire.
ConvS2S	Sekao ke porojeke ya go tlhola ditiro tsa go tlhola ditiro tse di ikaelelang go samagana le ntlha e e 40% ya batho mo baaging ba sa dire.
Setswana-to-Eng Speaker ConvS2S	Example is a project that will create jobs that will create jobs that aim at dealing with the point that that 40% of people in the society don't work.
Transformer	Sekao ke porojeke ya go tlhama ditiro tsa magae e e ikaelelang go samagana le ntlha ya gore 40% ya batho ba ba sa direng.
Setswana-to-Eng Speaker Transformer	Example is a rural job creation project that aims to deal with the issue that 40% of people are not working.

from additional language models or linguistic pre-processing, the architectures performed extremely competitively, with Transformer achieving a new state-of-the-art of 33.12 for English-to-Setswana translation.

4.2 Qualitative

Table 2 shows qualitative results for specific sentences from our test set. In order to understand the feasibility of using such models, a Setswana speaker translated the English-to-Setswana translations generated by our model back to English. Although not perfect, the translations capture much of the meaning from the original sentence. Impressively, the translations also use synonyms for other concepts: for example, Transformer translated "40% of people in a community are unemployed" to the Setswana equivalent of "40% of people are not working".

In conjunction with the quantitative results, these results confirm our hypothesis that the use of NMT systems, in particular the Transformer model, can improve the state of the art in English to Setswana translation.

5 Conclusion

Due to the rising need for African translations of online educational resources, the development of accurate machine translation systems for low-resourced languages has become an issue of importance.

We showed that state-of-the-art NMT architectures can significantly outperform existing SMT architectures for translation from English to Setswana with minimal hyperparameter optimization, and only a small amount of training time. This result suggests the promise of using the Transformer architecture to train models to translate other African languages. Future work includes training the Transformer architecture on multiple African languages at once, and then specialising on a specific language, as is done for Romanian by Gu *et al* [13].

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

6 Acknowledgements

We would like to thank the organisers of the Deep Learning Indaba. Without the Indaba we would never have met, nor would we have had the resources and confidence to pursue and submit such research. Thank you to Guy Bosa for aiding us with our qualitative translations.

References

[1] Nicholas Spaull. South africa's education crisis: The quality of education in south africa 1994-2011. *Johannesburg: Centre for Development and Enterprise*, pages 1–65, 2013.

[2] Mihir Parikh and Sameer Verma. Utilizing internet technologies to support learning: an empirical analysis. *International Journal of Information Management*, 22(1):27–46, 2002.

[3] W3Techs.com. Usage of content languages for websites. `https://w3techs.com/technologies/overview/content_language/all`, 2018.

[4] SA Stats. Census 2011 statistical release. *Pretoria, South Africa: Statistics South Africa*, 2012.

[5] Joyce Chepkemoi. What languages are spoken in rwanda? `https://www.worldatlas.com/articles/what-languages-are-spoken-in-rwanda.html`, Jul 2017.

[6] Ethiopia language stats. `http://www.nationmaster.com/country-info/profiles/Ethiopia/Language`, 2018.

[7] Leonard Muaka. Language perceptions and identity among kenyan speakers. In *Selected Proceeding of the 40th Annual Conference on African Linguistics*, pages 217–230. Cascadilla Proceedings Project Somerville, MA, 2011.

[8] Jan-Thorsten Peter, Tamer Alkhouli, Andreas Guta, and Hermann Ney. The rwth aachen university english-romanian machine translation system for wmt 2016. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 356–361, 2016.

[9] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

[10] Kato Ronald and Etienne Barnard. Statistical translation with scarce resources: a south african case study. 2006.

[11] Ilana Wilken, Marissa Griesel, and Cindy McKellar. Developing and improving a statistical machine translation system for english to setswana: a linguistically-motivated approach. In *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, page 114, 2012.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[13] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*, 2018.

[14] Cindy McKellar. Autshumato english-setswana parallel corpora. `https://rma.nwu.ac.za/index.php/resource-catalogue/autshumato-english-setswana-multi-bilingual-corpus.html`, 2018.

[15] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018.