MACHINE TRANSLATION BASELINES FOR ARABIC - SWAHILI

Anonymous authors
Paper under double-blind review

Abstract

Building neural machine translation (NMT) systems for low-resource languages poses several challenges, mainly due to the lack of parallel data. In this research, we propose a baseline NMT system for translating between Arabic and Swahili. Despite being spoken by nearly 300 million individuals worldwide, the parallel corpus between these two languages is severely underrepresented. To address this, we scraped and processed the largest high-quality parallel corpus of Swahili and Arabic to our knowledge. We then used state-of-the-art NMT models, including Transformers and multilingual variants of Transformers, to build a baseline for bidirectional Arabic-Swahili NMT. Finally, we report an increase in the performance of our NMT system using the back-translation technique.

1 Introduction

This paper presents a baseline study for translation between Swahili and Arabic. We utilize the latest state-of-the-art transformer-based models, setting a solid foundation for future research in this domain.

Our experiments show that transfer learning through fine-tuning larger and richer models is effective in improving the translation quality of low-resource language pairs. Specifically, we explore different pre-trained multilingual models such as M2M100 (Fan et al., 2020) and MT5 (Xue et al., 2020) and evaluate their performance for the task of Swahili-Arabic translation. We report our results in terms of BLEU (Papineni et al., 2002) scores. Our best-performing model achieved a BLEU score of 24.47 on the Arabic-Swahili test set, which is a significant improvement over our baseline Transformer model.

1.1 Our Contribution

1. We release the largest dataset for the Arabic-Swahili language pair collected by processing the Bible and aligning the sentences at the verse level.
2. We validate that using a small dataset to fine-tune a large model improves performance significantly for African languages even for a morphologically rich language like Arabic.

2 Related Work

Some work has been done on developing neural machine translation baselines for African languages (Ahia & Ogueji, 2020). However, there is no known work for this language pair. And we intend to lay the groundwork for future research.

3 METHODOLOGY

Three baseline models were trained using the Transformer architecture of Vaswani et al. (2017). we adopted Fairseq (Ott et al., 2019) to train an encoder-decoder transformer model from scratch utilizing sub-word tokenization (BPE) (Sennrich et al., 2015) method

to tackle the limited vocabulary issue. the configuration is based on Vaswani et al. (2017) default parameters with a vocabulary size of 12000. and it was trained for 450 epochs.
we conduct fine-tuning experiments with the pre-trained models : mT5 (Xue et al., 2020) with a vocabulary size of 128k and M2M100 (Fan et al., 2020) using Fairseq (Ott et al., 2019) and Hugging-face (Wolf et al., 2020) platform.
We trained both models with the same parameters to allow for a fair comparison. We use back-translation (Abdulmumin et al., 2021) with the aid of a monolingual corpus to examine its effect on our best-performing model.

3.1 DATASET

The construction of a parallel corpus is very challenging and requires extensive human expertise. We mapped each verse from the Holy Bible1 [1] in Arabic to its corresponding verse in Swahili. In doing so, we obtained over 30,000 parallel sentences. We ran a sanity check, and the blue score was 98.4, indicating that the data is aligned. We include the NTREX (Federmann et al., 2022) (News Test Reference for Evaluating MT Models) data set, which is a 2000-sentence parallel corpus.

4 Results

4.1 Quantitative

All models were trained using BPE (Sennrich et al., 2015). And the fine-tuned models outperform the baseline transformer significantly.

The table below shows our results obtained after conducting several experiments.

Table 1: Test Blue Scores

Direction	Transformer-base	M2M100	mt5	Back-Translation
Arabic to Swahili	10.51	24.47	15.13	
Swahili to Arabic	11.59	10.63	7.64	

4.2 Qualitative

The translation qualities were rated very well. in particular, Arabic to Swahili. Translation examples for our best-performing model are in the Appendix.

5 CONCLUSION AND FUTURE WORK

In conclusion, this paper has presented a baseline for low-resource machine translation between Arabic and Swahili. Despite the challenges of working with a low-resource language pair, we have shown that it is possible to achieve reasonable translation quality with the right combination of pre-processing techniques and neural machine translation models. We hope that our results will inspire further research in this area and help to improve access to information and services for Arabic and Swahili speakers, particularly in the context of the African Union's efforts to promote linguistic diversity and inclusivity.

In the future, we plan to collect more parallel data and experiment with different denoising techniques to improve the performance of our models. Also, we want to discover the effect of tweaking the base transformer architecture to better fit the purpose of low-resource translation.

[1] http://bible.org

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

References

Idris Abdulmumin, Bashir Shehu Galadanci, and Abubakar Isa. Enhanced back-translation for low resource neural machine translation using self-training. In Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers 3, pp. 355–371. Springer, 2021.

Orevaoghene Ahia and Kelechi Ogueji. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. arXiv preprint arXiv:2003.12660, 2020.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In Proceedings of the First Workshop on Scaling Up Multilingual Evaluation, pp. 21–24, Online, nov 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sumeval-1.4.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038, 2019.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

A Appendix

Below are some translation examples produced by our fine-tuned M2M model.

Table 2: Examples

Description	text
Reference	naye bwana akaniambia, nimewaona watu hawa nao ni watu wenye shingo ngumu sana
Model Translation	kisha bwana akaniambia, nimewaona watu hawa, nao watu hawa ni watu wenye shingo ngumu
Reference	hapo mwanzo alikuwako neno, huyo neno alikuwa pamoja na mungu, naye neno alikuwa mungu.
Model Translation	hapo mwanzo neno lilikuwepo, neno lilikuwepo pamoja na mungu na neno lilikuwepo pamoja na mungu.
Reference	yesu akajibu, rudini mkamwambie yohana yale mnayosikia na kuyaona
Model Translation	yesu akawajibu, nendeni mkamwambie yohana yale mliyoyasikia na kuyaona
Reference	hesabu ya waume wote wenye umri wa mwaka mmoja au zaidi waliohesabiwa walikuwa 6,200.
Model Translation	idadi ya wanaume wote wenye umri wa mwezi mmoja au zaidi walikuwa 6,200