The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation

Tadesse Destaw Belay[1], **Atnafu Lambebo Tonja**[2], **Olga Kolesnikova**[2], **Seid Muhie Yimam** [3], **Abinew Ali Ayele**[4], **Silesh Bogale Haile**[5], **Grigori Sidorov**[2], **and Alexander Gelbukh**[2]

[1] College of Informatics, Wollo University, Kombolcha, Ethiopia
tadesseit@gmail.com
[2]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico
alabedot2022@cic.ipn.mx, kolesolga@gmail.com, {lastname}@cic.ipn.mx
[3]Dept. of Informatics, Universität Hamburg, Hamburg, Germany
seid.muhie.yimam@uni-hamburg.de
[4]Faculty of Computing, Bahir Dar University, Bahir dar, Ethiopia
abinewaliayele@gmail.com
[5]Dept. of Computer Science, Assosa University, Assosa, Ethiopia
sileshibogale123@gmail.com

Abstract

Machine Translation (MT) is a sub-field of NLP that investigates how to use computer software to automatically translate text or speech from one language to another without human involvement. Nowadays, using deep neural networks for MT tasks has received a great deal of attention. This paper presents the first relatively large-scale Amharic-English parallel sentence dataset. Using these compiled data, we build bi-directional Amharic-English translation models by fine-tuning Facebook's multilingual pre-trained model achieving a BLEU score of 37.79 in Amharic-English translation and 32.74 in English-Amharic translation. Additionally, we explore the effects of Amharic homophone normalization on the MT task. The results show that normalization of Amharic homophone characters increases the performance of Amharic-English MT in both directions.

1 Introduction

Machine translation (MT) is one of the prominent tasks in NLP that is tackled in several ways (Bojar et al., 2017). It has a very long research history and passed through four stages, namely, Rule-based MT (Forcada et al., 2011), Statistical MT (SMT) (Koehn et al., 2007), hybrid MT, and Neural MT (NMT) (Cho et al., 2014; Kalchbrenner and Blunsom, 2013).

Recently, using neural networks for MT tasks has received great attention. NMT also improves the training process due to the end-to-end procedure without tedious feature engineering and complex setups. Transformer models with the pre-training approach is a new NMT strategy entirely based on attention mechanisms proposed in 2017 (Vaswani et al., 2017; Raganato et al., 2018). It has become the state-of-the-art model for many artificial intelligence tasks, including MT.

The focus of MT research for Amharic[1] has been on rule-based and SMT methods. In this work, we used the transformer model as a baseline translation system to explore the applicability of the Facebook multilingual pre-trained model (Fan et al., 2020).

The main contributions of this work are: 1) exploration of the Amharic-English and English-Amharic MT tasks, 2) introduction of the first large-scale publicly available Amharic-English translation parallel dataset, 3) development and implementation of state-of-the-art Amharic-English translation models, and 4) investigation of the effect of Amharic homophone character normalization on the MT task.

2 Methodology

Related Works: Related works for the Amharic-English and English-Amharic translation tasks are summarized in Table 1.

Building Parallel Dataset: As MT requires parallel documents as an input, Table 2 shows the potential Amharic-English bi-lingual resources. As shown in the table, the total number of collected parallel sentences is around 1.1M, while the unique

[1]Amharic is the official language of the Federal Democratic Republic of Ethiopia (FDRE) and for many regional states in the country.

Table 1: Amharic-English and English-Amharic MT studies in terms of dataset size, method(s), and BLEU score

Authors	Trans. direction	# of Sentences	Method(s)	BLEU score
Biadgligne and Smaïli (2021)	En→Am	225, 304	SMT,NMT	26.47, 32.44
Gezmu et al. (2021)	Am→En	45,364	PSMT, NMT	20.2, 26.6
Abate et al. (2018)	Am→En, En→Am	40,726	SMT	22.68, 13.31
Teshome et al. (2015)	En→Am	18,432	PSMT	37.5
Teshome and Besacier (2012)	En→Am	18,432	PSMT	35.32
Ashengo et al. (2021)	En→Am	8,603	CBMT with RNN	11.34
Ambaye and Yared (2000)	En→Am	37,970	SMT	18.74
Hadgu et al. (2020)	Am→En	977	Google, Yandex	23.2, 4.8
	En→Am	1915	Google, Yandex	9.6, 1.3

parallel sentences are 888k. This is due to duplication in the sources we used. This unique parallel sentence is the largest to date then. In addition to the available datasets in Table 2, we have contributed to the MT research field by creating a new parallel corpus with 33,955 sentence pairs extracted from such news platforms as Ethiopian Press Agency[2], Fana Broadcasting Corporate[3], and Walta Information Center[4]. As the data we used is from different sources, it includes various domains such as religious, politics, economics, sports, news, among others.

We performed a series of data pre-processing: data cleaning, abbreviation expansion, Latin character lowercase, duplicated sentence removal, and Amharic homophone character normalization.

Table 2: Amharic-English parallel data sources

Am→En MT data sources	# of sentences
Am-En ELRA-W0074	13,347
Biadgligne and Smaïli (2021)	225,304
Horn MT	2,030
Am-En MT corpus	53,312
Gezmu et al. (2021)	145,364
Abate et al. (2018)	40,726
Lison and Tiedemann (2016)	562,141
Tracey and Strassel (2020)	60,884*
Admasethiopia	153
MT Evaluation Dataset	2,914
Newly curated (our data)	**33,955**
Total	1,140,130
Unique sentence pairs	888,837

*Not available freely

[2]https://www.press.et/
[3]https://www.fanabc.com/
[4]https://waltainfo.com/

Amharic Homophone Character: In Amharic writing, there are different characters with the same sound which are called homophones. Homophones with different symbols in Amharic text might have different writing standards and different meanings. However, they are also considered redundant alphabets by most of the users, especially by the online and social media communities.

The current trend in Amharic NLP research is to normalize the homophone characters into a single representation (Woldeyohannis and Meshesha, 2017; Abate et al., 2018; Gezmu et al., 2021; Biadgligne and Smaïli, 2021). There is no study to show that normalization has a positive or negative impact on MT tasks. We have developed our translation models in the regular (unnormalized) and the normalized forms (representing different homophones as a single character) to analyze the impact of normalization. For the normalized one, we made normalization on the entire training and testing data.

Proposed MT Models: We trained Transformer sequence-to-sequence models from scratch and we used the pre-trained multi-lingual model (M2M100 418M) proposed by Facebook Fan et al. (2020) for our bi-directional MT experiments.

3 Experimental Setup and Results

We used Google Colab Pro+ to train our bi-directional Amharic to English translation models. During model training, the parallel sentences were divided into 80% for training, 10% validation, and 10% for testing. The automatic evaluation was made using BLEU metric (Papineni et al., 2002).

As shown in Table 3, the multilingual pre-trained-based model outperforms the Transformer-based models in both translation directions. The

Table 3: Experimental results

Models & Trans. direction	Regular	Normalized
Transformer Am→En	14.78	16.26
Transformer En→Am	10.79	13.06
M2M100 48M Am→En	34.12	**37.79**
M2M100 48M En→Am	29.65	**32.74**

score of the Amharic-English or vice versa from pre-trained model that we have adopted (M2M100) is not mentioned in the original work. When we compare our results with the attempts mentioned in related works, Table 1, our research shows an improvement in parallel corpus size and BLEU scores using the multilingual translation model. The baseline transformer based models scores low even from the traditional approaches in the related work. This might be due to: we have used the default hyper-parameters during model training and the two approaches are different. Our results also clearly show the effect of homophone character normalization in the performance of bi-directional Amharic-English translation. For both (Transformer and pre-trained) models, homophone character normalization increased the performance of the NMT system.

4 Conclusion and Future Work

In this paper, we presented our Amharic-English parallel corpus and bi-directional MT experiments. We gathered more than 888K parallel unique sentences, applied different preprocessing techniques, and built state-of-the-art Amharic-English MT models. This is the first large-scale parallel data that can be used as a good benchmark for future MT research. The resources such as MT datasets and the models are available publicly in GitHub repository[5]. We will expand this work to more languages by including other Ethiopian low-resourced languages.

[5] https://github.com/uhh-lt/ethiopicmodels

References

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tadesse Ambaye and Mekuria Yared. 2000. English to Amharic machine translation using statistical machine translation. *Master's thesis*.

Yeabsira Asefa Ashengo, Rosa Tsegaye Aga, and Surafel Lemma Abebe. 2021. Context based machine translation with recurrent neural network for English–Amharic translation. *Machine Translation*, 35(1):19–36.

Yohanens Biadgligne and Kamel Smaïli. 2021. Parallel corpora preparation for English-Amharic machine translation. In *International Work-Conference on Artificial Neural Networks*, pages 443–455. Springer, Cham.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

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. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for Amharic-English machine translation. *arXiv preprint arXiv:2104.03543*.

Asmelash Teka Hadgu, Adam Beaudoin, and Abel Aregawi. 2020. Evaluating amharic machine translation. *arXiv preprint arXiv:2003.14386*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. pages 923–929.

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

Alessandro Raganato, Jörg Tiedemann, et al. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297. The Association for Computational Linguistics.

Mulu Gebreegziabher Teshome and Laurent Besacier. 2012. Preliminary experiments on English-Amharic statistical machine translation. In *Spoken Language Technologies for Under-Resourced Languages*, pages 36–41, Cape Town, South Africa.

Mulu Gebreegziabher Teshome, Laurent Besacier, Girma Taye, and Dereje Teferi. 2015. Phoneme-based English-Amharic statistical machine translation. In *AFRICON 2015*, pages 1–5, Addis Ababa, Ethiopia. IEEE.

Jennifer Tracey and Stephanie Strassel. 2020. Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 277–284, Marseille, France. European Language Resources association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Michael Melese Woldeyohannis and Million Meshesha. 2017. Experimenting statistical machine translation for Ethiopic Semitic languages: The case of Amharic-Tigrigna. In *International Conference on Information and Communication Technology for Development for Africa (ICT4DA 2017)*, volume 244, pages 140–149. Springer, Cham.