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

Tadesse Destaw Belay
College of Informatics
Wollo University
Kombolcha, Ethiopia
tadesseit@gmail.com

Atnafu Lambebo Tonja
Centro de Investigación en Computación
Instituto Politécnico Nacional
Mexico City, Mexico
alabedot2022@cic.ipn.mx

Olga Kolesnikova
Centro de Investigación en Computación
Instituto Politécnico Nacional
Mexico City, Mexico
kolesolga@gmail.com

Seid Muhie Yimam
Dept. of Informatics
Universität Hamburg
Hamburg, Germany
seid.muhie.yimam@uni-hamburg.de

Abinew Ali Ayele
ICT4D Research Center
Bahir Dar University
Bahir dar, Ethiopia
abinewaliayele@gmail.com

Silesh Bogale Haile
Dept. of Computer Science
Assosa University
Assosa, Ethiopia
sileshibogale123@gmail.com

Grigori Sidorov
Centro de Investigación en Computación
Instituto Politécnico Nacional
Mexico City, Mexico
sidorov@cic.ipn.mx

Alexander Gelbukh
Centro de Investigación en Computación
Instituto Politécnico Nacional
Mexico City, Mexico
gelbukh@cic.ipn.mx

***Abstract*—Machine translation (MT) is one of the prominent tasks in natural language processing whose objective is to translate texts automatically from one natural language to another. Nowadays, using deep neural networks for MT task has received a great deal of attention. These networks require lots of data to learn abstract representations of the input and store it in continuous vectors. 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 the existing Facebook M2M100 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 machine translation task. The results show that normalization of Amharic homophone characters increases the performance of Amharic-English machine translation in both directions.**

***Index Terms*—Neural machine translation, pre-trained models, Amharic-English MT, homophone normalization, low-resourced language**

I. Introduction

Machine translation (MT) is a sub-field of natural language processing (NLP) that investigates how to use computer software to automatically translate text or speech from one language to another without human involvement. MT is one of the prominent tasks in NLP that is tackled in several ways [1]. The first MT research began at about 1950s and in 1952 the first International Conference on Machine Translation was organized at the Massachusetts Institute of Technology (MIT). It has long research history and experienced four stages, namely, Rule-based MT [2], Statistical MT (SMT) [3], hybrid MT, and Neural MT (NMT) [4], [5].

The most severe drawback of the rule-based method is that it has ignored the need for context information in the translation process. It is highly dependent on hand-crafted features. Phrase-based SMT (PBSMT), the most prevalent version of SMT, generates translation by segmenting the source sentence into several phrases and performing phrase translation and replacement. It may ignore the long sentence dependency and require high computing devices [6]. Recently, using deep neural networks for MT task has received great attention. NMT also improves training procedures due to the end-to-end procedure without tedious feature engineering and complex setups. NMT employs such techniques as recurrent neural network (RNN) [7], convolutional neural network (CNN) [8], and self-attention network (Transformer) [9].

Transformer models with the pre-training approach is a new NMT strategy entirely based on attention mechanisms proposed in 2017 [9]. Among the different neural network architectures, the Transformer model has emerged as the dominant NMT paradigm [10]–[12]. It has become the state-of-the-art model for many artificial intelligence tasks, including machine translation. In terms of model, the Transformer-based pre-trained models are fast to fine-tune, highly accurate and has been proven to outperform widely used recurrent networks [6], [13], [14].

The focus of MT research for the Amharic language 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 Facebook M2M100 multi-lingual pre-trained language model for Amharic-English translation in both directions. Furthermore, this research work investigated the impact of normalization of the Amharic homophones on Amharic-English MT tasks.

The main contributions of this work are:

1) Exploration of the Amharic-English and English-Amharic machine translation 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.
4) Investigation of the effect of Amharic homophone character normalization on the machine translation task.

The rest of this paper is organized as follows. Section II presents a detail description of Amharic language while Section III shows the motivation for this research. In Section IV, we review related work. Section V describes the existing parallel corpus and the collection of a new corpus from the news domain. The general pre-processing steps applied to both corpora are presented in Section VI. Section VII discusses the proposed NMT models and Section VIII gives the experimental results. In the end, Section IX concludes the paper and sheds some light on possible future work.

II. Amharic Language

Amharic is the second most spoken Semitic language next to Arabic which has its own alphabet and writing scripts called 'Fidel', that was borrowed from Ge'ez, another Ethiopian Semitic language. Fidel is a syllable-based writing system where the consonants and vowels co-exist within each graphic symbol. The Amharic language is spoken by more than 57 million people with up to 32 million native speakers and 25 million non-native speakers [15]. Amharic is the working language of the Federal Democratic Republic of Ethiopia (FDRE) and for many regional states in the country. In Amharic, there are 34 core characters each having seven different derivatives to represent vowels. In addition, it has 20 labialized characters, more than 20 numerals, and 8 punctuation marks. Amharic uses a total of more than 310 characters. The language is known for being morphologically complex and it is highly inflectional. Unlike English, French, Spanish, Japanese, and Chinese, Amharic is considered low-resource because the data are not well organized and technologically less supported [16].

III. Motivation

Nowadays advancement in technology has made the lifestyle of human beings much easier by helping daily activities. One of the applications that solved communication barriers between people speaking different languages is machine translation. Many big technology companies such as Google, Microsoft, IBM, etc. provide translation services for many languages to facilitate communication between people without using a human translator. However, the quality of NMT is massively dependent on quantity, quality, and relevance of the training dataset [17]. Such companies have achieved promising results for bilingual high-resource languages, but they are inadequate for low-resource languages like Amharic.

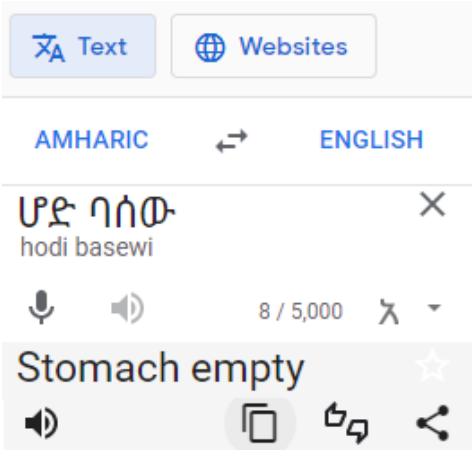

Fig. 1. Examples of Google Amharic to English translation

Figure 1 shows Google translation of Amharic words into English. Google translated the Amharic input as "stomach empty" which is a wrong translation. The correct literal meaning will be "he become Disappointed". This shows that the applications used by companies that provide translation system like Google require improvements. One of the cause for poor performance of MT systems for languages like Amharic is availability of limited resource in digital space [17]. In this research work we present a newly curated English-Amharic parallel dataset that can be used for MT research and help to solve the performance issue of an Amharic MT system.

On the other hand, Amharic is one of morphologically rich language and normalization of the Amharic homophone characters might have an impact on such downstream NLP applications as MT and sentiment analysis [18]. This research work is intended to study the effect of homophone normalization on Amharic-English machine translation. Furthermore, expanding the translation dataset and developing state-of-the-art bi-directional Amharic-English translation models are our motivations to carryout this research work.

IV. Related Work

Many automatic translation works have been carried out for the major pairs of European and Asian languages, taking advantage of large-scale parallel corpora. However, very few research studies have been conducted on low-resource languages like Amharic to English due to its scarcity of parallel data. In this section, we have focused on exploring how machine translation is conducted for the Amharic language. Among recent works, Biadgligne and Smaïli [19] described the development of an English-Amharic Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) experiments, achieving 26.47 and 32.44 BLEU scores, respectively. They harvested and used 225,304 parallel sentences in different domains. To the best of our knowledge, Biadgligne and Smaïli [19] work is the largest dataset used in Amharic machine translation research work.

Gezmu et al. [20] trained neural machine translation and phrase-based statistical machine translation models using 145,364 parallel English-Amharic sentences, achieved a BLEU score of 20.2 and 26.6 respectively. They concluded that neural machine translation models outperform phrase-based statistical machine translation models. Abate et al. [21] focused on the development of parallel corpora for English and five Ethiopian languages, including Amharic. They used 40,726 parallel sentences for Amharic and bi-directional Statistical Machine Translation (SMT) technique. Finally, they achieved a BLEU score 22.68 in Amharic-English translation.

Teshome et al. [22] considered the application of phone-based statistical machine translation for English-Amharic translation. They used 18,432 parallel sentences and achieved a BLEU score of 37.5, a gain of 2.21 BLEU point from their own previous baseline phrase-based MT experiments. Further more, Tracey and Strassel [23] developed a low-resource language dataset for emergent incidents (LORELEI). Their data includes multiple languages and was published in the Linguistic Data Consortium (LDC) where Amharic is among the languages with 60,884 Amharic-English parallel sentences. However, this dataset is not freely available for further experiments.

The remain and addressed related works are summarized in Table I, with dataset and methods used. As we can see in this related work table, only a few studies have been conducted for the translation from Amharic into English or vice versa and most of them were conducted using traditional approaches with a small number of parallel sentences. This is due to unavailability of enough Amharic linguistic resources for deep learning experiments.

V. Building a Parallel Dataset

In this section, we have identified available data sources that we have used for the research work. As machine translation requires parallel documents as an input, Table II shows the potential Amharic-English bi-lingual resources. It can be seen in the table that the largest parallel corpus for Amharic-English language pairs was collected by Biadgligne and Smaïli [28]. In addition to the available datasets in Table II, we have contributed to the MT research field by creating a new parallel corpus with 33,955 sentence pairs extracted text from such news platforms as Ethiopian Press Agency[1], Fana Broadcasting Corporate[2], and Walta Information Center[3]. As the data we used is from different sources, it includes various domains such as religious (Bible and Quran), politics, economics, sports, news, among others.

As one can see in Table II, the total number of parallel sentences is about 1.1M, while the unique parallel sentences are 888,837. This is due to duplication in the sources we used. This unique parallel sentence is the largest to date then.

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

VI. Data pre-processing

Before performing experiments, the development of every NLP task begins with a text pre-processing step [29]. As our data were collected from different sources, we noticed a lot of text irregularities. Some data was too noisy so we eliminated it from our corpus. Then we performed a series of pre-processing steps to canonize all tokens in Amharic and English sentences. In these steps, the following tasks were performed: data cleaning (removing emojis, URL), abbreviation expansion, Latin character lowercase, duplicated sentence removal, and Amharic homophone character normalization. For abbreviation expansion, we created a list of known abbreviations for both Amharic and English, then expanded them to their full writing form. Most of English abbreviations were collected from GitHub repositories[4].

TABLE II
Available Amharic-English parallel data sources

Data source	# Sentence pairs	Accessible
Am-En ELRA-W0074[5]	13,347	yes
Biadgligne and Smaïli [19]	225,304	yes
Horn MT[6]	2,030	yes
Am-En MT corpus[7]	53,312	yes
Gezmu et al. [20]	145,364	yes
Abate et al. [21]	40,726	yes
Open Parallel Corpus (OPUS) [30]	562,141	yes
LORELEI-Amharic [23]	60,884	no
Admasethiopia[8]	153	yes
MT Evaluation Dataset[9]	2,914	yes
Newly curated (our data)	**33,955**	**yes**
Total	1,140,130	yes
Unique sentence pairs	888,837	yes

Homophone character normalization: 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 as redundant alphabets by most of the users, specially by the online and social media communities.

The current trend in Amharic NLP research is to normalize the homophone characters into a single representation [16], [19]–[21]. Belay et al. [18] have studied the impact of normalization for some downstreamm NLP applications. They showed that homophone normalization improves the performance of some downstream applications such as information retrieval (IR) but not recommended for sentiment analysis. However, there is no study to show that normalization helps to build efficient machine translation models or not. We 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

[4]https://github.com/JRC1995/Machine-Translation-Transformers
[5]http://catalog.elra.info/en-us/repository/browse/ELRA-W0074/
[6]https://github.com/asmelashteka/HornMT
[7]https://github.com/adtsegaye/Amharic-English-Machine-Translation-Corpus
[8]https://github.com/admasethiopia/parallel-text
[9]https://doi.org/10.5281/zenodo.3734260

TABLE I
AMHARIC-ENGLISH AND ENGLISH-AMHARIC MT STUDIES IN TERMS OF DATASET SIZE, METHOD(S) USED, AND BLEU SCORE ACHIEVED

Authors	Trans. direction	# Dataset used	Method(s)	BLEU score
Biadgligne and Smaïli [19]	En→Am	225,304	Statistical machine translation	26.47
			Neural machine translation	32.44
Gezmu et al. [20]	Am→En	45,364	Phrase-based statistical machine translation	20.2
			Neural Machine Translation	26.6
Abate et al. [21]	Am→En	40,726	Statistical machine translation	22.68
	En→Am		Statistical machine translation	13.31
Teshome et al. [22]	En→Am	18,432	Phoneme-based statistical machine translation	37.5
Teshome and Besacier [24]	En→Am	18,432	Phrase-based statistical machine translation	35.32
Ashengo et al. [25]	En→Am	8,603	Combination of context-based MT (CBMT) with RNN	11.34
Tadesse and Mekuria. [26]	En→Am	37,970	Statistical machine translation	18.74
Hadgu et al. [27]	Am→En	977	Google translate, Yandex translate	23.2, 4.8
	En→Am	1915	Google translate, Yandex translate	9.6, 1.3

the normalized model, we used frequency-based homophone normalizer [18] replacing the set of characters with similar function with a single most frequently used character.

1st	2nd	3rd	4th	5th	6th	7th
ሀ (ha)	ሁ (hu)	ሂ (hi)	ሃ (hā)	ሄ (hé)	ህ (he/h)	ሆ (ho)
ሐ (ḥa)	ሑ (ḥu)	ሒ (ḥi)	ሓ (ḥā)	ሔ (ḥé)	ሕ (ḥe/ḥ)	ሖ (ḥo)
ኀ (h̲a)	ኁ (h̲u)	ኂ (h̲i)	ኃ (h̲ā)	ኄ (h̲é)	ኅ (h̲e/h̲)	ኆ (h̲o)
ኸ (xa)	ኹ (xu)	ኺ (xi)	ኻ (xā)	ኼ (xé)	ኽ (xe/x)	ኾ (xo)
አ ('a)	ኡ ('u)	ኢ ('i)	ኣ ('ā)	ኤ ('é)	እ ('e)	ኦ ('o)
ዐ ('a)	ዑ ('u)	ዒ ('i)	ዓ ('ā)	ዔ ('é)	ዕ ('e)	ዖ ('o)
ሰ (se)	ሱ (su)	ሲ (si)	ሳ (sā)	ሴ (sé)	ስ (se/s)	ሶ (so)
ሠ (śa)	ሡ (śu)	ሢ (śi)	ሣ (śā)	ሤ (śé)	ሥ (śe/ś)	ሦ (śo)
ጸ (ṣa)	ጹ (ṣu)	ጺ (ṣi)	ጻ (ṣā)	ጼ (ṣé)	ጽ (ṣe/ṣ)	ጾ (ṣo)
ፀ (ṡa)	ፁ (ṡu)	ፂ (ṡi)	ፃ (ṡā)	ፄ (ṡé)	ፅ (ṡe/ṡ)	ፆ (ṡo)

Fig. 2. Amharic homophone characters with their 7 derivatives using their International Phonetic Alphabet (IPA) notation. The normalized homophones is the first one in boldface in cell of each row except the of 4th column of the 1st and 2nd row where the 1st character was used as the normalized one.

VII. THE PROPOSED NEURAL MACHINE TRANSLATION MODELS

In this section, we describe the transformer-based pre-trained models that could be used to train Neural Machine Translation systems. Unlike other neural networks such as RNNs, the Transformer does not necessarily process the input data in sequential order. Instead, the self-attention mechanism identifies the context which gives meaning to each position in the input sequence, allowing more parallelization, and reducing the training time. The architecture of the Transformer network follows the so called encoder-decoder paradigm, trained in an end-to-end fashion. The encoder is used to represent the source sentence as a semantic vector, while the decoder takes in the semantic vector and makes prediction to a target sentence. The Transformer model, which applies a self-attention approach to measure the strength of a relationship between two words in a sentence (contextual information), has contributed to improving performance in MT and various natural language processing tasks.

A. Pre-trained language models (PLMs)

Pre-trained Language Models (PLMs) are large neural networks used in a wide variety of NLP tasks [31]. Models are first pre-trained over a large text corpus and then fine-tuned on a downstream task. PLMs are thought of as good language encoders, supplying basic language understanding capabilities that can be used with ease for many downstream tasks [31], [32].

Fan et al. [33], Facebook researchers, proposed a multilingual encoder-decoder model trained for Many-to-Many multilingual translation with 418M parameters (M2M100 418M). This multilingual machine translation model is based on the Transformer sequence-to-sequence architecture. The model can directly translate between the 9,900 directions of 100 languages. We used this pre-trained model for our bi-directional translation experiments. This pre-trained multi-lingual model is available at Hugging Face[10]

VIII. EXPERIMENTAL SETUP AND RESULTS

A. Experimental setup

To demonstrate the effectiveness of our approach, we built our baseline Transformer models and fine-tuned the available pre-trained model. We used Google colab pro+ to train our bi-directional Amharic to English translation models. During model training, the parallel sentences for Amharic and English were divided into 80% for training, 10% validation, and 10% for testing. Automatic evaluation was made using Bilingual Evaluation Understudy (BLEU) metric [34]. BLEU score is defined in the range between 0 and 1, where 1 is a perfect match with the reference and 0 is for no words matched.

Transformer baseline: We trained Transformer sequence-to-sequence models from scratch for bi-directional Amharic-English NMT using OpenNMT with TensorFlow deep learning framework [35]. We tokenized the text using Byte Pair Encoding (BPE) [36] subword tokenization, which is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data. The parameters used to train the model are 512 hidden units, 6 layers, a learning rate of 0.0001, and

[10] https://huggingface.co/facebook/m2m100_418M

a maximum step of 50K, a batch size of 32, and the Adam optimizer.

Pre-trained model: We used the multilingual Facebook M2M-100 pre-trained model with 418M parameters [33] to fine-tune into bi-directional Amharic-English NMT. For fine-tuning, we used max source & target length of 128 per device train and validation, a batch size of 4, and 3 epochs.

B. Results and discussion

In our experiments, we built Transformer models and fine-tuned M2M100_418M, the multi-lingual pre-trained language model for bi-directional Amharic-English translation. All models were built on a regular (unnormalized) and normalized datasets to study the effect of Amharic homophone normalization in Amharic-English translation. Table III shows the experimental results of Transformer and PLMs of M2M100 in both directions.

TABLE III
EXPERIMENTAL RESULTS

Models	Regular	Normalized
Transformer (Amharic→English)	14.78	16.26
Transformer (English→Amharic)	10.79	13.06
M2M100 48M (Amharic→English)	34.12	**37.79**
M2M100 48M (English→Amharic)	29.65	**32.74**

As it can be observed from the results in Table III, the multilingual translation model M2M-100 outperforms the Transformer-based models in both directions. When we compare our results with the attempts mentioned in Section IV, our research shows an improvement in parallel corpus size and BLEU scores using multilingual translation model. We can evaluate our models in the translation directions as follows.

Amharic-English Translation: For both Amharic-English and English-Amharic translations, the Transformer models did not work well, performed even less than some traditional methods. The pre-trained transformer-based model performed better with 34.12 and 37.79 BLEU score for Amharic-English translation, on the regular and normalized data, respectively. This clearly shows that pre-trained based translation outperforms the baseline system.

English-Amharic Translation: Among the works conducted for English-Amharic translation, the work by Teshome et al. [22] reached a BLEU score of 37.53 using phoneme-based statistical MT, while our pre-trained model showed a BLEU score of 29.65 and 32.74 on regular and normalized data, respectively. This phoneme-based work [22] is built by converting the Amharic syllables to phoneme-based characters; however 1) the conversion is rule-based (unable to handle all exceptions), 2) Amharic homophones do not have constant Latin representations, and 3) we used new train and test sets. So, the results are not comparable. Another insight from this work is pre-trained model still needs improvement to translate from technologically favored languages to languages like Amharic which is morphologically rich.

Our results also clearly show the effect of homophone character normalization in the performance of bi-directional

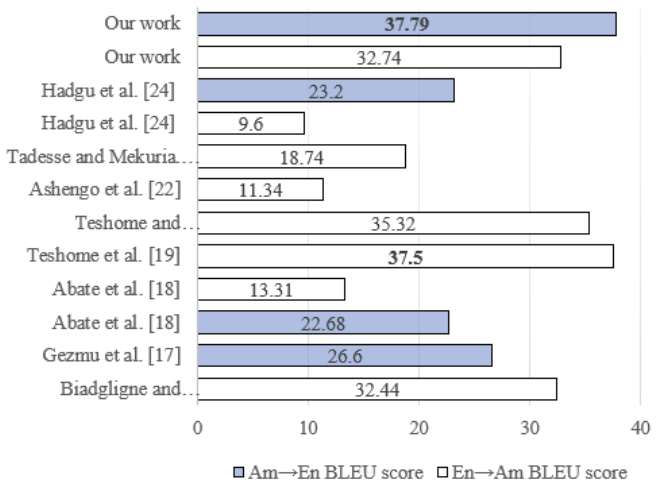

Fig. 3. Reporting of our work result with other previous works

Amharic-English translation. For both (Transformer and pre-trained) models, homophone character normalization increased the performance of NMT system. Accordingly, comparing our results with the previous attempts as shown in Figure 3, we can confirm that our models demonstrated a 0.29% BLEU score increase. This BLEU score improvement was made using our new training and testing sets. This is because, first, all the mentioned MT data is not available; second, the available data is not split into constant training and testing categories for future evaluation purposes. For the next time, our dataset will be available in train, test, and validation files.

IX. CONCLUSION AND FUTURE WORK

In this paper, we presented our Amharic-English parallel corpus and bi-directional machine translation experiments. We gathered more than 888K parallel unique sentences and applied different pre-processing techniques. This is the first large-scale parallel data (more than 5 times bigger than the data in the previously conducted MT works) which can be used as a good benchmark for future machine translation research. The main gaps from the previous work were that they used small data and were not ready for future comparison (benchmark) by splitting the data into constant train, test, and validation sets. Our work will solve this issue and be used as a benchmark. According to our result, we can conclude that the pre-trained models outperformed the baseline Transformer-based model. The morphological richness of the Amharic language and the size of the parallel dataset has a great impact on Amharic-English MT experiments. For the future, we will expand this work for more languages by including other Ethiopian low-resourced languages and also use data augmentation techniques. Additionally, we plan to explore the applicability of other pre-trained language models for Amharic-English translations. Our dataset of parallel Amharic-English sentence pairs, the models, and pre-processing scripts

will be released in the GitHub repository[11]

REFERENCES

[1] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva *et al.*, "Findings of the 2017 conference on machine translation (wmt17)," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 169–214.

[2] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers, "Apertium: a free/open-source platform for rule-based machine translation," *Machine translation*, vol. 25, no. 2, pp. 127–144, 2011.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens *et al.*, "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*. Association for Computational Linguistics, 2007, pp. 177–180.

[4] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[5] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 conference on empirical methods in natural language processing*. Seattle, Washington, USA: Association for Computational Linguistics, 2013, pp. 1700–1709.

[6] L. H. Baniata, I. Ampomah, S. Park *et al.*, "A transformer-based neural machine translation model for Arabic dialects that utilizes subword units," *Sensors*, vol. 21, no. 19, pp. 1–28, 2021.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] J. Gehring, M. Auli, D. Grangier, and Y. N. Dauphin, "A convolutional encoder model for neural machine translation," *arXiv preprint arXiv:1611.02344*, 2016.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.

[10] A. Raganato, J. Tiedemann *et al.*, "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*. The Association for Computational Linguistics, 2018, pp. 287–297. [Online]. Available: http://hdl.handle.net/10138/263704

[11] M. G. Yigezu, M. M. Woldeyohannis, and A. L. Tonja, "Multilingual neural machine translation for low resourced languages: Ometo-english," in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*. IEEE, 2021, pp. 89–94.

[12] A. L. Tonja, M. M. Woldeyohannis, and M. G. Yigezu, "A parallel corpora for bi-directional neural machine translation for low resourced ethiopian languages," in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, 2021, pp. 71–76.

[13] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the transformer translation model with document-level context," *arXiv preprint arXiv:1810.03581*, 2018.

[14] S. Yang, Y. Wang, and X. Chu, "A survey of deep learning techniques for neural machine translation," *arXiv preprint arXiv:2002.07526*, 2020.

[15] D. M. Eberhard, G. F. Simons, and C. D. Fennig, "Ethnologue: Languages of the world (2022)," 2022. [Online]. Available: URL:https://www.ethnologue.com/

[16] M. M. Woldeyohannis and M. Meshesha, "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)*, vol. 244. Springer, Cham, 2017, pp. 140–149.

[17] B. Ahmadnia and B. J. Dorr, "Augmenting neural machine translation through round-trip training approach," *Open Computer Science*, vol. 9, no. 1, pp. 268–278, 2019.

[18] T. D. Belay, A. A. Ayele, G. Gelaye, S. M. Yimam, and C. Biemann, "Impacts of homophone normalization on semantic models for amharic," in *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA 2021)*. Bahir Dar, Ethiopia: IEEE, 2021, pp. 101–106.

[19] Y. Biadgligne and K. Smaïli, "Parallel corpora preparation for English-Amharic machine translation," in *International Work-Conference on Artificial Neural Networks*. Springer, Cham, 2021, pp. 443–455.

[20] A. M. Gezmu, A. Nürnberger, and T. B. Bati, "Extended parallel corpus for Amharic-English machine translation," *arXiv preprint arXiv:2104.03543*, 2021.

[21] S. T. Abate, M. Melese, M. Y. Tachbelie, M. Meshesha, S. Atinafu, W. Mulugeta, Y. Assabie, H. Abera, B. Ephrem, T. Abebe, W. Tsegaye, A. Lemma, T. Andargie, and S. Shifaw, "Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, aug 2018, pp. 3102–3111. [Online]. Available: https://aclanthology.org/C18-1262

[22] M. G. Teshome, L. Besacier, G. Taye, and D. Teferi, "Phoneme-based English-Amharic statistical machine translation," in *AFRICON 2015*. Addis Ababa, Ethiopia: IEEE, 2015, pp. 1–5.

[23] J. Tracey and S. Strassel, "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)*. Marseille, France: European Language Resources association, 2020, pp. 277–284. [Online]. Available: https://aclanthology.org/2020.sltu-1.39

[24] M. G. Teshome and L. Besacier, "Preliminary experiments on English-Amharic statistical machine translation," in *Spoken Language Technologies for Under-Resourced Languages*, Cape Town, South Africa, 2012, pp. 36–41. [Online]. Available: https://www.isca-speech.org/archive/

[25] Y. A. Ashengo, R. T. Aga, and S. L. Abebe, "Context based machine translation with recurrent neural network for English–Amharic translation," *Machine Translation*, vol. 35, no. 1, pp. 19–36, 2021.

[26] T. Ambaye and M. Yared, "English to Amharic machine translation using statistical machine translation," *Master's thesis*, 2000.

[27] A. T. Hadgu, A. Beaudoin, and A. Aregawi, "Evaluating amharic machine translation," *arXiv preprint arXiv:2003.14386*, 2020.

[28] y. Biadgligne and K. Smaïli, "Offline Corpus Augmentation for English-Amharic Machine Translation," in *2022 The 5th International Conference on Information and Computer Technologies*, New York, United States, Mar. 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03547539

[29] A. A. Kumar and S. Chandrasekhar, "Text data pre-processing and dimensionality reduction techniques for document clustering," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 5, pp. 1–6, 2012.

[30] P. Lison and J. Tiedemann, "Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles," pp. 923–929, 2016.

[31] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and improving consistency in pretrained language models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1012–1031, 2021.

[32] R. Weng, H. Yu, S. Huang, S. Cheng, and W. Luo, "Acquiring knowledge from pre-trained model to neural machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9266–9273.

[33] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," *arXiv preprint*, 2020.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics, 2002, pp. 311–318.

[35] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," *arXiv preprint arXiv:1701.02810*, 2017.

[36] P. Gage, "A new algorithm for data compression," *C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[11] https://github.com/atnafuatx/EthioNMT-datasets