EthioMT: Parallel Corpus for Low-resource Ethiopian Languages

Atnafu Lambebo Tonja [♠,◆,*], **Olga Kolesnikova** [♠],
Alexander Gelbukh [♠], **Jugal Kalita** [♣],

[♠] Instituto Politécnico Nacional, Mexico, [◆] Lelapa AI,
[♣] University of Colorado Colorado Springs, USA

Abstract

Recent research in natural language processing (NLP) has achieved impressive performance in tasks such as machine translation (MT), news classification, and question-answering in high-resource languages. However, the performance of MT leaves much to be desired for low-resource languages. This is due to the smaller size of available parallel corpora in these languages, if such corpora are available at all. NLP in Ethiopian languages suffers from the same issues due to the unavailability of publicly accessible datasets for NLP tasks, including MT. To help the research community and foster research for Ethiopian languages, we introduce EthioMT – a new parallel corpus for 15 languages. We also create a new benchmark by collecting a dataset for better-researched languages in Ethiopia. We evaluate the newly collected corpus and the benchmark dataset for 23 Ethiopian languages using transformer and fine-tuning approaches.

Keywords: Parallel corpus, EthioMT, Machine Translation, low resource language, Ethiopian languages

1. Introduction

In recent years, due to advances in deep learning approaches such as the development of transformers (Vaswani et al., 2017), machine translation (MT), a core task in natural language processing (NLP), has shown dramatic improvements in terms of coverage and translation quality (Wang et al., 2021). It is well-known that a critical requirement for advancing MT is the availability of parallel corpora. The availability of parallel corpora is also necessary to facilitate the incorporation of languages in MT applications like Google Translation, Bing, and DeepL (Van der Meer, 2019). The majority of the languages in the world do not have access to such translation tools since only a few high-resource languages have received significant attention (Tonja et al., 2023b).

Most models and methods developed for high-resource languages do not work well in low-resource settings (Costa-jussà et al., 2022; Tonja et al., 2023b; King, 2015). Low-resource languages have also suffered from language technology designs (Joshi et al., 2019; Tonja et al., 2022). Creating powerful novel methods for language applications is challenging when resources are limited and only a small amount of even unlabeled data is available. The problem is exacerbated when no parallel dataset exists for specific languages (Joshi et al., 2020; Ranathunga et al., 2023; Adebara and Abdul-Mageed, 2022).

Ethiopia is a country that stands out for its remarkable cultural and linguistic diversity, with over 85 spoken languages (Woldemariam, 2007). Only a few languages of Ethiopia have received attention in the area of NLP research and application development. Most languages have been left behind due to resource limitation (Costa-jussà et al., 2022; Tonja et al., 2023b). It is hard to find publicly available datasets for Ethiopian languages to pursue NLP research because many researchers do not make their datasets publicly accessible (Tonja et al., 2023b). The unavailability of benchmark datasets and results for NLP tasks, including MT, makes research for newcomers and interested parties very difficult. This is obviously more difficult for languages with limited data in different digital forms.

This paper introduces EthioMT: a parallel corpus for low-resource Ethiopian languages paired with English, and a benchmark dataset and experimental results for 23 Ethiopian languages. Our contributions are the following: **(1)** We create a **new parallel corpus for 15 Ethiopian languages** paired with English. **(2)** We introduce the **first benchmark dataset and results for relatively better resourced Ethiopian** (Amharic, Afaan Oromo, Tigrinya and Somali) **languages**. **(3)** We evaluate MT performance with the **new corpus and present benchmark results. (4)** We **open-source** the parallel corpus to foster collaboration and facilitate research and development in low-resource Ethiopian languages.

2. Related work

Ethiopian languages are categorized as low-resource due to the unavailability of resources for NLP tasks, including MT (Tonja et al., 2023b). Although MT is a better-researched area for Ethiopian languages compared to other NLP applications (Tonja et al., 2023b), only a handful of languages have received adequate attention from researchers.

[*] Work done during an internship at the University of Colorado Colorado Springs.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 107–114
25 May, 2024.

Researched Languages Compared to other Ethiopian languages, the following languages have received significant attention from researchers. Nevertheless, the collected corpora are not found in one location. It is hard to find benchmark datasets in these languages and datasets and associated results to reproduce and compare MT approaches.

Amharic - Researchers have collected parallel datasets and proposed different MT approaches for Amharic-English translation (Kenny, 2018; Teshome and Besacier, 2012; Hadgu et al., 2020; Ashengo et al., 2021; Biadgligne and Smaïli, 2022; Belay et al., 2022; Gezmu et al., 2021b,a; Biadgligne and Smaïli, 2021).

Afaan Oromo - Similarly, there have been attempts to create Afaan Oromo-English MT datasets (Meshesha and Solomon, 2018; Solomon et al., 2017; Adugna and Eisele, 2010; Chala et al., 2021; Gemechu and Kanagachidambaresan, 2021).

Tigrinya - For Tigrinya-English MT, researchers have attempted to create parallel datasets (Tedla and Yamamoto, 2016, 2017; Berihu et al., 2020; Azath and Kiros, 2020; Kidane et al., 2021).

Multilingual MT Some researchers have included Ethiopian languages with other languages in multilingual MT systems. Lakew et al. (2020) collected and created benchmark results for five African languages, including those mentioned above from Ethiopia. Costa-jussà et al. (2022), Goyal et al. (2022) and Fan et al. (2021) included Ethiopian languages in their multilingual MT models and benchmark test sets. Vegi et al. (2022) crawled a multilingual parallel dataset for African languages, including Amharic and Afaan Oromo from Ethiopia.

Other languages There have been efforts to create and collect MT datasets for other Ethiopian languages. For example, Tonja et al. (2021) presented a parallel corpus for four low-resourced Ethiopian languages (Wolaita, Gamo, Gofa, and Dawuro).

3. EthioMT

3.1. Discussion of Languages

In this section, we enumerate languages included in the EthioMT corpus. Languages include in the EthioMT corpus belong to Afro-Asiatic and Nilo-Saharan language families.

3.1.1. Afro-Asiatic language family

The Afro-Asiatic language family comprises about 250 languages spoken in North Africa, parts of sub-Saharan Africa, and the Middle East. Languages belonging to this family are grouped into six sub-groups: Berber, Chadic, Cushitic, Egyptian, Omotic, and Semitic (Epstein and Kole, 1998). EthioMT contains languages belonging to the Omotic, Cushitic, and Semitic sub-groups.

1) Omotic Languages are a group of languages spoken in southwestern Ethiopia, in the Omo River region. The Ge'ez script is used to write some of the Omotic languages and the Latin script for others (Amha, 2017). Languages belonging to this group that we included in EthioMT are given below.

Basketo is spoken in the Basketo special woreda of the South Ethiopia Regional State. The Basketo language is also called Basketto, Baskatta, Mesketo, Misketto, and Basketo-Dokka. The speakers call the language "Masketo", while their neighbors call it "Basketo". The language has two dialects, Doko (Dokko) and Dollo (Dollo).

Dawuro is a language spoken by about 1.09 million people in the Dawro zone of the South West Ethiopia Peoples' Region. It is also known as Dauro, Dawragna, Dawrogna, Ometay, Cullo, or Kullo. The language has four dialects: Konta, Kucha, Longkhai, and Yawngkon.

Gamo is spoken by around 1.63 million people in the Gamo Zone of the South Ethiopia Regional State. The speakers call the language Gamotstso.

Gofa refers to the language spoken in the Gofa zone of the South Ethiopia Regional State with around 392,000 speakers.

Kafa, also known as Kefa or Kafi noono is a North Omotic language spoken in Ethiopia. It is spoken by about 830,000 people in the Keffa Zone in the South West Ethiopia Peoples' Region. The language is mainly spoken in and around the town of Bonga.

Male is spoken in the Omo Region of Ethiopia. The Male people maintain their language vigorously despite exposure to outside pressures and languages.

Shakicho, also known as Mocha, Shakacho, or Shekka, is spoken in the Sheka Zone of southwestern Ethiopia. It is closely related to Kafa. Loan words from Majang and Amharic influence the language's vocabulary.

Wolaytta is a North Omotic language spoken by the Welayta people in the Wolayita Zone of Ethiopia. It is estimated that 2 million people speak Wolaytta.

2) Cushitic languages are spoken primarily in the Horn of Africa, including Djibouti, Eritrea, Ethiopia, Somalia, and Kenya (Comrie, 2002). The Cushitic languages use the Latin and Ge'ez script. Languages belonging to this family that are included in the EthioMT group are discussed below.

Afar is spoken by the Afar people in Ethiopia, Eritrea, and Djibouti. It is also known as Afar Af, Afaraf, and Qafar af. About 1.5 million people speak Afar, the closest relative to the Saho language.

Afaan Oromo, also known as Oromo, is spoken by about 37 million people in Ethiopia, Kenya, Somalia, and Egypt. It is the third-largest language in Africa and the largest language in the Cushitic group in terms of speakers. The Oromo people are the largest ethnic group in Ethiopia and account for more than 40 percent of the population.

Language	Family	Explored prev.	No. of Speaker	Domain	Size
Afar (aar)	Afro-Asiatic / Cushitic	×	1.5M	Religious	11K
Afaan Oromo (orm)	Afro-Asiatic / Cushitic	✓	37M	Misc	**2.9M**
Awngi (awn)	Afro-Asiatic / Cushitic	×	490K	Religious	7K
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	✓	57M	Misc	**1.5M**
Basketo (bst)	Afro-Asiatic/ Omotic	×	93K	Religious	7K
Dawuro (dwr)	Afro-Asiatic/ Omotic	✓	1.5M	Religious	7K
Dashenech (dsh)	Afro-Asiatic/ Cushitic	×	99K	Religious	7K
Geez (gez)	Afro-Asiatic / Ethio-Semitic	×	UNK	Religious	7K
Gamo (gmv)	Afro-Asiatic / Omotic	✓	1.09M	Religious	7K
Gofa (gof)	Afro-Asiatic / Omotic	✓	392K	Religious	7K
Gurage (sgw)	Afro-Asiatic / Ethio-Semitic	×	5.8M	Religious	28K
Hadiya (hdy)	Afro-Asiatic / Cushitic	×	1.3M	Religious	28K
Kafa (kbr)	Afro-Asiatic / Omotic	×	830K	Religious	28K
Korate (kxc)	Afro-Asiatic / Cushitic	×	500K	Religious	7K
Majang (mpe)	Nilo-Saharan / Eastern Sudanic	×	66K	Religious	9K
Male (mdy)	Afro-Asiatic / Omotic	×	105K	Religious	7K
Murule (mur)	Nilo-Saharan / Eastern Sudanic	×	300K	Religious	9K
Nuer (nus)	Nilo-Saharan /Eastern Sudanic	×	900K	Religious	29K
Shakicho (moy)	Afro-Asiatic / Omotic	×	80K	Religious	7K
Sidama (sid)	Afro-Asiatic / Cushitic	×	4M	Religious	28K
Somali (som)	Afro-Asiatic / Cushitic	✓	22.3M	Misc	**1.2M**
Tigrinya (tir)	Afro-Asiatic / Ethio-Semitic	✓	9M	Misc	**140K**
Wolaytta (wal)	Afro-Asiatic / Omotic	✓	7M	Religious	29K

Table 1: Languages and dataset details for **EthioMT** corpus. It shows languages, language families, the number of speakers, the domain, and the size of the collected dataset. In domain column **Misc** indicates *mixed* corpus collected from religious, news, and other sources. **Bold and underlined** size indicates a dataset collected from different repositories and published works and merged into one dataset for the language to create a benchmark dataset

Awngi is a Central Cushitic language spoken by about 400,000 people in northwestern Ethiopia. It is also known as Awiya, Awi, Agaw, Agau, Agew, Agow, Awawar, and Damot. Most speakers live in the Agew Awi Zone of the Amhara Region. Awngi is an Afro-Asiatic language spoken in parts of the Metekel Zone of the Benishangul-Gumuz Region.

Dashenech is also known as Dasenech, Daasanech, or Daasanach. The Daasanach people speak it in Ethiopia, South Sudan, and Kenya. The Daasanach people primarily live in the Lower Omo Valley of southwestern Ethiopia, along the eastern shore of Lake Turkana in Kenya, and in some parts of South Sudan.

Hadiya is spoken by the Hadiya people of Ethiopia. The language is also known as Hadiyyisa, Hadiyigna, Adiya, Adea, Adiye, Hadia, Hadiya, and Hadya. It is a Highland East Cushitic language. The Hadiya people are an ancient indigenous group in the southern part of Ethiopia. There are 1.4 million speakers of the Hadiya language, with 1.25 million of them speaking it as their mother tongue.

Korate is a Lowland East Cushitic language spoken by the Konso people in southwest Ethiopia. It has approximately 500,000 native speakers. The language has five dialects: Duuro, Fasha, Karatti, Kholme, and Komso. The two main dialects are Fasha and Karatti. Konso is closely related to Dirasha (also known as Gidole). It is used as a "trade language" or lingua franca beyond the area of the Konso people. The Konso people are a Cushitic ethnic group who live in large towns in south-central Ethiopia.

Sidama, or Sidaamu Afoo, is a Cushitic language spoken by the Sidama people in southern Ethiopia. It uses the Latin alphabet. Almost nine million people speak Sidama. It is the official language of the Sidama National Regional State (SNRS) and is used as a medium of instruction in primary schools. Sidama is a branch of the Highland East Cushitic family.

Somali is the official language of Somalia, spoken by 6.5 million people. It is also spoken in Ethiopia, Djibouti, and Kenya. The total number of speakers worldwide is estimated at nearly 22 million. Its closest relative is the Oromo language, spoken in parts of Ethiopia and Kenya. Other related languages include Afar and Saho.

3) Semitic languages belong to a subfamily of the Afro-Asiatic language family, including Hebrew, Aramaic, Arabic, and Ethiopic. Most scripts used to write Semitic languages are abjad. Abjad refers to an alphabetic script that omits some or all vowels. Languages belonging to this group that we study are given below.

Amharic is spoken by the Amhara and other regions in Ethiopia. It is the second most-spoken Semitic language in the world, after Arabic. Amharic is the official language of Ethiopia and has been since the 14th century. It is also spoken in other countries, including Eritrea, Canada, the United States, and Sweden. Amharic is written using graphemes called *fidal*, which means "script", "alphabet", "letter", or "character".

Ge'ez is an ancient Semitic language that originated in Eritrea and northern Ethiopia. Ge'ez is believed to be around 5,000 years old, making it older than Hebrew and other Northern Semitic languages. Orthodox and Catholic churches in Eritrea and Ethiopia still use it as a liturgical language. Ge'ez went extinct as a natural language over 1,000 years ago. It was written in two systems: an abjad and later an abugida.

Gurage is spoken by the Gurage people in central Ethiopia. The Gurage languages are written using the Ge'ez script, which is also used for other Ethiopian languages. The Gurage languages are not always mutually intelligible.

Tigrinya is spoken by about 9 million people, primarily in Eritrea and Ethiopia. It is written in the Ge'ez script, which is also used for Amharic, but the grammar and usage of Tigrinya differs significantly from Amharic.

3.1.2. Nilo-Saharan language family

Nilo-Saharan languages are a group of languages that form one of the four language families on the African continent (Dimmendaal et al., 2019). The family covers major areas east and north of Lake Victoria in East Africa and extends westward to the Niger Valley in Mali, West Africa (Comrie, 2002). Nilo-Saharan constitutes ten distinct and separate language families, including Eastern Sudanic.

Eastern Sudanic languages are a group of ten families of languages that constitute a branch of the Nilo-Saharan language family. Eastern Sudanic languages are spoken from southern Egypt to northern Tanzania. The languages used in our study by this group are given below.

Majang is spoken by the Majangir people of Ethiopia. It is a member of the Surmic language cluster, but it is the most isolated one in the group. It is classified as part of the Eastern Sudanic branch of the Nilo-Saharan language family. The Majang people live in scattered settlements in southwestern Ethiopia. They live around the urban areas of Tepi and Mett'i, southwest of Mizan Teferi and towards Gambela.

Murle is spoken by the Murle people in South Sudan and Ethiopia. The language is also known as Ajibba, Beir, Merule, Mourle, and Murule. The Murle language is part of the Surmic language family and has three dialects: Lotilla, Boma, and Olam. The Murle people number between 300,000 and 400,000. They live in Pibor County in the southeastern Upper Nile (Jonglei)

Nuer or Thok Naath is a West Nilotic language spoken by the Nuer people of South Sudan and western Ethiopia. The language is written in a Latin-based alphabet, similar to Dinka and Atuot. Over 900,000 people speak the Nuer language in diaspora communities in East Africa, Australia, and the USA.

4. Dataset

4.1. Dataset Collection

We collected datasets for 16 languages from religious domains from a website[1]. In addition to that, for Amharic, Afaan Oromo, Somali, and Tigrinya, we collected publicly available datasets (Abate et al., 2019; Lakew et al., 2020; Vegi et al., 2022) from different domains to create one benchmark dataset per language. For Dawuro, Gamo, Gofa, and Wolaita languages, we used Tonja et al. (2021) dataset to create benchmark results for fine-tuned models. A web crawler was used for each article to extract the Bible data from websites after identifying the structure of web documents. Python libraries such as requests, regular expression (RE), and Beautiful Soup (BS) were utilized to analyze website structure and extract article content from a given URL.

4.2. Sentence Alignment

After collecting the corpus for the languages, we aligned each sentence of the Ethiopian languages to a sentence in English data to prepare the dataset for the MT experiment. We followed the same procedure as Tonja et al. (2023a) to perform sentence alignment.

4.3. Dataset Pre-processing

After aligning the texts of the Ethiopian languages with their equivalent translations in English, we pre-processed the corpus before splitting it for our experiments. The pre-processing steps included removing the numeric and special character symbols, etc. We also removed parallel sentences that contain less than five words. For the baseline experiments, we split the pre-processed corpus into training, development, and test sets in the ratio of 70:10:20, respectively. Table 1 shows detailed information on selected languages, language families, domain, and their dataset size.

[1] https://www.bible.com/

5. Baseline Models

We used the following two approaches to evaluate the newly collected corpus's usability and our new benchmark dataset of four (amh, orm, som, and tir) Ethiopian languages.

The baseline transformer is a type of neural network architecture first introduced in the paper *Attention Is All You Need* (Vaswani et al., 2017). The key innovation of the Transformer architecture is the attention mechanism, which allows the network to selectively focus on different parts of the input sequence when making predictions. This contrasts traditional recurrent neural networks (RNNs), which process input sequentially and are prone to the vanishing gradient problem.

In the transformer architecture, multiple self-attention layers and feed-forward neural networks process elements of the input sequence in parallel. Each layer can be considered a "block" that takes the previous layer's output as input and applies its transformations to it. The self-attention mechanism allows the network to weigh the importance of each element in the input sequence when making predictions. In contrast, the feed-forward networks help to capture non-linear relationships among the components.

Transformers are state-of-the-art approaches widely used in NLP tasks such as MT, text summarization, and sentiment analysis. Table 3 shows parameters set up for the transformer model.

Parameters	Values
encoder_layer	6
encoder_attention_head	4
decoder_layer	6
batch_size	512
batch_type	token
decoder_attention_head	8
hidden_size	256
embed_dim	256
dropout	0.2
beam_size	5
optimizer	adam
tokenizer_type	sentencepiece
max_input_length	150

Table 2: Parameters used for transformer training

Fine tuning is the process of using a pre-trained MT model and adapting it to a specific translation task, such as translating between a particular language pair or in a specific domain. The process of fine-tuning involves taking the pre-trained model, which has already learned representations of words and phrases from a large corpus of text, and training it on a smaller dataset of specific task examples. This involves updating the pre-trained model's parameters to better capture the patterns and structures in the target translation task.

Fine-tuning can be helpful in MT because it allows the pre-trained model to quickly adapt to a new task without having to train a new model from scratch. This is especially beneficial when working with limited data or when there is a need to quickly adapt to changing translation requirements. We used **M2M100-48** a multilingual encoder-decoder (seq-to-seq) model trained for many-to-many multilingual translation (Fan et al., 2021). We used a model with 48M parameters due to computing resource limitations. We used the following parameters to fine-tune the m2m100 model.

Parameters	Values
encoder_layer	12
encoder_attention_head	16
decoder_layer	12
batch_size	512
batch_type	token
decoder_attention_head	16
hidden_size	4096
embed_dim	1024
attention_dropout	0.1
beam_size	5

Table 3: Parameters used for m2m100-48 fine-tuning

6. Results and Discussions

We evaluated the above approaches in bidirectional translation from Ethiopian languages to English and From English to Ethiopian languages. We used Sacrebleu (Post, 2018) evaluation metrics to evaluate translation models. Tables 4 and 5 show the translation results in both directions.

6.1. Using English as a source language

Table 4 shows the translation results from English to Ethiopian languages. When comparing the results of the two approaches, we observe poor performance when using a transformer rather than fine-tuning the m2m100 model. As we can see from the result, the performance of the transformer model also varies in the ranges of 0.01 – 17.8 spBLEU from language to language with different corpus sizes. This shows that a bilingual translation model trained from scratch performs poorly for low-resource language training compared to other approaches like fine-tuning multilingual translation models. Fine-tuning the multilingual model shows better results than the model built from scratch for English to Ethiopian language translation. In the fine-tuning approach, we can also observe a clear score difference between languages with larger corpora (amh, orm, tir, som) and others (e.g awn, aar, bst, etc.). This shows that fine-tuning the multilingual model will work well for languages with the largest (e.g. orm, amh) corpus sizes than languages with

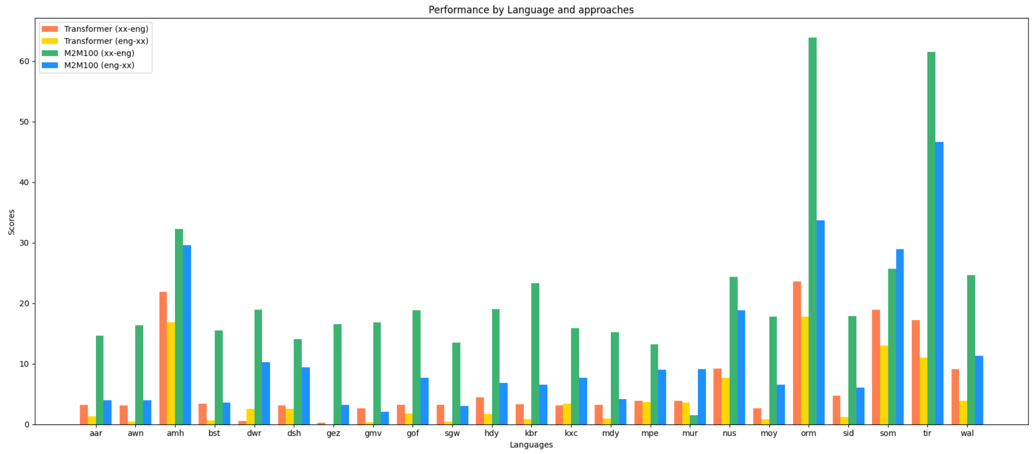

Figure 1: Benchmark translation results for transformer and fine-tuned approaches in both (from and to English/Ethiopian languages) direction

Model	en-xx																							
	aar	**awn**	**amh**	**bst**	**dwr**	**dsh**	**gez**	**gmv**	**gof**	**sgw**	**hdy**	**kbr**	**kxc**	**mdy**	**mpe**	**mur**	**nus**	**moy**	**orm**	**sid**	**som**	**tir**	**wal**	**Avg.**
	Bleu Score																							
Transformer	1.28	0.41	16.79	0.6	2.57	2.51	0.01	0.34	1.82	0.41	1.69	0.87	3.36	0.90	3.65	3.58	7.73	0.87	17.8	1.19	13.06	11.07	3.84	4.18
m2m100-fine-tuned	3.95	3.93	29.63	3.61	10.23	9.45	3.25	2.03	7.65	3.04	6.80	6.58	7.69	4.15	9.03	9.10	18.79	6.58	33.7	6.10	28.9	46.63	11.32	11.83

Table 4: Benchmark translation results from English to Ethiopian languages

Model	xx-en																							
	aar	**awn**	**amh**	**bst**	**dwr**	**dsh**	**gez**	**gmv**	**gof**	**sgw**	**hdy**	**kbr**	**kxc**	**mdy**	**mpe**	**mur**	**nus**	**moy**	**orm**	**sid**	**som**	**tir**	**wal**	**Avg.**
	Bleu Score																							
Transformer	3.18	3.14	21.9	3.39	0.52	3.07	0.28	2.68	3.21	3.18	4.42	3.26	3.14	3.21	3.91	3.92	9.23	2.63	23.6	4.77	18.9	17.2	9.16	6.60
m2m100-fine-tuned	15.61	16.32	65.34	15.47	18.92	14.11	16.57	16.79	18.79	13.52	19.04	23.27	15.90	15.20	13.26	1.48	24.40	17.78	63.9	17.86	25.71	61.50	24.62	21.79

Table 5: Benchmark translation results from Ethiopian languages to English

small (e.g. awn, bst, etc.) corpus sizes. We can also see from the results that both approaches work well for languages with mixed-domain texts compared to one domain (religion).

6.2. Using English as a target language

Table 5 shows the translation result when using English as a target language. Similarly, as we can see from the results, the transformer model performs poorly compared to the fine-tuned model when translating from Ethiopian languages to English. Compared to Table 4, translating to English shows improvements in the transformer model for similar languages. We observe that the fine-tuned model shows better Bleu scores when translating to English than when translating to Ethiopian languages. The results show that languages with large datasets have the highest performance. This shows that both models show improvements when translating from Ethiopian to English, while when translating from English to Ethiopian languages, the model is struggling with translation.

7. Conclusion and Future Works

This paper presents EthioMT, a new MT corpus for low-resource Ethiopian languages paired with English, and discusses MT experiments with results. We also present a new benchmark dataset for four Ethiopian languages collected from public repositories. We obtained benchmark results with new train, validation, and test set splits and evaluated the new corpus and new benchmark dataset using a transformer and fine-tuning multilingual translation models. From the two approaches, fine-tuning of the multilingual model outperformed the transformer approach in both translation directions.

In the future, we will work to increase the corpus sizes of the low-resource languages by extracting text from scanned documents and different sources. In addition, we will evaluate different MT approaches to low-resource languages to improve performance.

8. Bibliographical References

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

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021b. Neural machine translation for amharic-english translation. In *ICAART (1)*, pages 526–532.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

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.

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

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023a. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. *arXiv preprint arXiv:2305.17404*.

Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, and Mesay Gemeda Yigezu. 2021. 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)*, pages 71–76. IEEE.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Mexican International Conference on Artificial Intelligence*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023b. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *arXiv preprint arXiv:2303.14406*.

Azeb Amha. 2017. The omotic language family. Cambridge University Press.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages*. Ph.D. thesis.

Bernard Comrie. 2002. Languages of the world: who speaks what. In *An encyclopedia of language*, pages 529–543. Routledge.

Dorothy Kenny. 2018. Machine translation. In *The Routledge handbook of translation and philosophy*, pages 428–445. Routledge.

Ebisa A Gemechu and GR Kanagachidambaresan. 2021. Machine learning approach to english-afaan oromo text-text translation: Using attention based neural machine translation. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 80–85. IEEE.

Edmund L Epstein and Robert Kole. 1998. *The language of African literature*. Africa World Press.

Gerrit J Dimmendaal, Colleen Ahland, Angelika Jakobi, and Constance Kutsch Lojenga. 2019. Linguistic features and typologies in languages commonly referred to as 'nilo-saharan'. *Cambridge Handbook of African Languages*, pages 326–381.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*.

Hirut Woldemariam. 2007. The challenges of mother-tongue education in ethiopia: The case of north omo area. *Language Matters*, 38(2):210–235.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. 2021. An exploration of data augmentation techniques for improving english to tigrinya translation. *arXiv preprint arXiv:2103.16789*.

M Azath and Tsegay Kiros. 2020. Statistical machine translator for english to tigrigna translation. *Int. J. Sci. Technol. Res*, 9(1):2095–2099.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Million Meshesha and Yitayew Solomon. 2018. English-afaan oromo statistical machine translation. *International Journal of Computational Linguistic (IJCL)*, 9(1).

Mulu Gebreegziabher Teshome and Laurent Besacier. 2012. Preliminary experiments on english-amharic statistical machine translation. In *Spoken Language Technologies for Under-Resourced Languages*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Pavanpankaj Vegi, J Sivabhavani, Biswajit Paul, Abhinav Mishra, Prashant Banjare, KR Prasanna, and Chitra Viswanathan. 2022. Webcrawl african: A multilingual parallel corpora for african languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Sisay Adugna and Andreas Eisele. 2010. English—oromo machine translation: An experiment using a statistical approach. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Sisay Chala, Bekele Debisa, Amante Diriba, Silas Getachew, Chala Getu, and Solomon Shiferaw. 2021. Crowdsourcing parallel corpus for english-oromo neural machine translation using community engagement platform. *arXiv preprint arXiv:2102.07539*.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Biniyam Ephrem, Tewodros Gebreselassie, et al. 2019. English-ethiopian languages statistical machine translation. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 27–30.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *arXiv preprint arXiv:2003.14402*.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. 2022. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 84–89. IEEE.

Jaap Van der Meer. 2019. Translation technology–past, present and future. *The Bloomsbury companion to language industry studies*, pages 285–310.

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.

Yemane Tedla and Kazuhide Yamamoto. 2016. The effect of shallow segmentation on english-tigrinya statistical machine translation. In *2016 International Conference on Asian Language Processing (IALP)*, pages 79–82. IEEE.

Yemane Tedla and Kazuhide Yamamoto. 2017. Morphological segmentation for english-to-tigrinya statistical machinetranslation. *Int. J. Asian Lang. Process*, 27(2):95–110.

Yitayew Solomon, Million Meshesha, and Wendewesen Endale. 2017. Optimal alignment for bidirectional afaan oromo-english statistical machine translation. *vol*, 3:73–77.

Yohanens Biadgligne and Kamel Smaïli. 2021. Parallel corpora preparation for english-amharic machine translation. In *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part I 16*, pages 443–455. Springer.

Yohannes Biadgligne and Kamel Smaïli. 2022. Offline corpus augmentation for english-amharic machine translation. In *2022 5th International Conference on Information and Computer Technologies (ICICT)*, pages 128–135. IEEE.

Zemicheal Berihu, Gebremariam Mesfin Assres, Mulugeta Atsbaha, and Tor-Morten Grønli. 2020. Enhancing bi-directional english-tigrigna machine translation using hybrid approach. In *Norsk IKT-konferanse for forskning og utdanning*, 1.