AmhEn: Amharic-English Large Parallel Corpus for Machine Translation

Atnafu Lambebo Tonja (alambedot2022@cic.ipn.mx)
Instituto Politécnico Nacional

Tadesse Destaw Belay (tadesseit@gmail.com)
Wollo University

Olga Kolesnikova (kolesolga@gmail.com)
Instituto Politécnico Nacional

Seid Muhie Yimam (seid.muhie.yimam@uni-hamburg.de)
Universität Hamburg

Abinew Ali Ayele (abinewaliayele@gmail.com)
Bahir Dar University

Grigori Sidorov (sidorov@cic.ipn.mx)
Instituto Politécnico Nacional

Alexander Gelbukh (gelbukh@cic.ipn.mx)
Instituto Politécnico Nacional

Research Article

Keywords:

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

AmhEn: Amharic-English Large Parallel Corpus for Machine Translation

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

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico.
[2]College of Informatics, Wollo University, Kombolcha, Ethiopia.
[3]Dept. of Informatics, Universitat Hamburg, Hamburg , Germany.
[4]ICT4D Research Center, Bahir Dar University, Bahir Dar, Ethiopia .

*Corresponding author(s). E-mail(s): alambedot2022@cic.ipn.mx;
Contributing authors: tadesseit@gmail.com;
†These authors contributed equally to this work.

Abstract

Recently, using deep neural networks for machine translation (MT) tasks has received great attention. In order for these networks to learn abstract representations of the input and store them as continuous vectors, they need a lot of data. However, very few research studies have been conducted on low-resource languages like Amharic. The progress of an Amharic-English machine translation task in both directions is affected by the lack of clean, easy-to-find, and up-to-date parallel language corpora. This paper presents the first relatively large-scale Amharic-English parallel corpora (above 1.1 million) for machine translation tasks. We ran experiments with recurrent neural networks (RNN) and Transformer in various hyper-parameter settings to investigate the usability of our dataset. Additionally, we explore the effects of Amharic homophone character normalization on machine translation. We have released the dataset in both unnormalized and normalized forms. Our dataset is available in train, test, and validation split files.

Keywords: machine translation, MT dataset, Amharic-English MT, low-resource machine translation, Amharic parallel corpus

1 Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence that employs computational techniques for the purpose of learning, understanding, and producing human-language content (Hirschberg and Manning, 2015). Machine Translation (MT) is one of the widely used NLP applications that carry out the automatic translation from one language to another in order to facilitate communication between people who speak different languages. The goal of MT is to automatically translate texts or speech from one natural language to another without human involvement. For machine translation, researchers use many methodologies, including rule-based MT (Forcada et al, 2011), statistical MT (Koehn et al, 2007), hybrid MT, and neural MT (NMT) (Cho et al, 2014; Kalchbrenner and Blunsom, 2013). The current state-of-the-art neural MT trained on enormous datasets including sentences in a source language and their corresponding target language translations, is the most effective of these systems. Recently, using deep neural networks for MT tasks has received great attention, but the performance of NMT as a data-driven approach massively depends on the quantity, quality, and relevance of the training dataset (Tonja et al, 2021, 2022; Yigezu et al, 2021). The end-to-end process of NMT, which does not require tedious feature engineering or complicated setups, also makes training better. NMT employs such techniques as recurrent neural network (RNN) (Bahdanau et al, 2014), convolutional neural network (CNN) (Gehring et al, 2016), and self-attention network (Transformer) (Vaswani et al, 2017).

The progress of research in the area of machine translation for Amharic-English is slowed by the lack of clean, large, easily accessible, and up-to-date parallel language corpora. As a result, there have been a few attempts in machine translation for Amharic to English or vice versa using small corpora, and most of the research have been done by using the traditional machine translation approaches (Teshome and Besacier, 2012; Ashengo et al, 2021; Teshome et al, 2015). However, the datasets applied in these works are still not readily available to the research community, this hinders their efforts to work on Amharic-English translation. This paper presents the first relatively large-scale Amharic-English parallel dataset for machine translation. Additionally, we explore the effects of Amharic homophone character normalization on the machine translation task.

The rest of the paper is organized as follows: Section 2 describes the Amharic language, Section 3 describes previous research related to Amharic-English machine translation, Section 4 describes the data collection framework, Section 5 explains the baseline experiments and their results, Section 6 concludes the paper.

2 Amharic language

Amharic(āmarinya) is a Semitic language related to Hebrew, Arabic, and Syriac, with the second-highest number of speakers after Arabic. It has its own

alphabet and writing script called Fidel(fīdeli) or Ethiopic script, which was adopted from the Ge'ez script. Ge'ez is another ancient Ethiopian Semitic language. Amharic is the official working language of the Federal Democratic Republic of Ethiopia (FDRE) and of many regional states in the country. It is also used by the government, public media and mass communication (like TV, radio, books, entertainment, etc.), and national commerce. The Amharic language is spoken by more than 57 million people, with up to 32 million native speakers and 25 million non-native speakers (Eberhard et al, 2022). Language-specific characteristics are described in the following sections.

2.1 Amharic Morphology

In Amharic, the root word is a set of consonants that bear the basic meaning of the lexical item, whereas a pattern is composed of a set of vowels inserted between the consonants of the root. These vowel patterns, together with affixes (prefixes, infixes, and suffixes), result in derived words. Such derivational processes make this language more morphologically complex. Additionally, an orthographic word may attach some syntactic words like prepositions, conjunctions, negations, etc., which make word forms highly varied (Gasser, 2011). Furthermore, nominals are inflected for number, gender, definiteness, and case, whereas verbs are inflected for person, number, gender, tense, aspect, and mood. So, in cases like these and others, Amharic is among the morphologically rich Semitic languages.

2.2 Writing System

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 writing system is syllabary, where each character represents a consonant and a vowel. The basic features of such writing system are that each character gets its basic shape from the consonant of the syllable, and the vowel is represented through systematic modifications of the basic shapes. Its writing and readings are from left to right. The Amharic alphabets with their full seven derivatives are listed in Appendix A1.

3 Related Work

Many automatic translation works have been carried out for the major pairs of English, Central European, and Asian languages, taking advantage of large-scale parallel corpora. Due to the lack of parallel data, however, not many studies have been done on low-resource languages like Amharic-English translation. In this section, we have focuse on exploring existing works on Amharic machine translation, their summaries are presented in Table 1, along with the dataset and methods used. As we can see in the table, only a few studies have been conducted for the translation of Amharic into English or vice

versa, and most of them were done using traditional approaches with a small number of parallel sentences. This is due to the unavailability of enough Amharic-to-English linguistic resources for deep-learning MT experiments.

4 Building parallel dataset

The general framework for developing the dataset is shown in Figure 1. As shown in Figure 1, it had two main tasks. The first task was data collection, which defined the sources for collecting the parallel Amharic and English sentences. After collecting parallel sentences, the second task was determining the data pre-processing strategies. The details of data collection and pre-processing tasks are described in the following subsections, Sections 4.1 and 4.2, respectively.

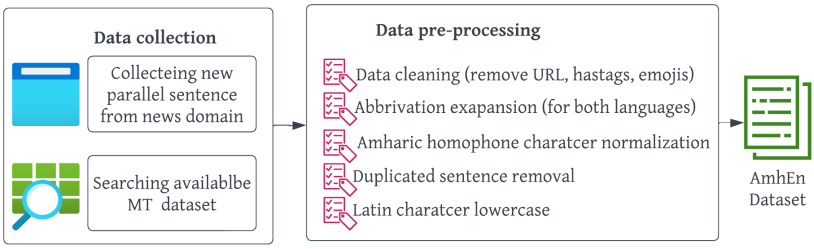

Fig. 1 Amharic-English MT data collection and pre-processing pipelines

4.1 Data collection

As MT requires parallel sentences of source and target languages as input, Table 2 shows the existing Amharic-English bi-lingual parallel datasets. As it can be seen in the table, the largest parallel corpus for Amharic-English language pairs was collected from the Open Parallel Corpus (OPUS) (Lison and Tiedemann, 2016). In addition to the available dataset sources 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[1], Fana Broadcasting Corporate[2], and Walta Information Center[3]. As the data were collected from different sources, they included various domains such as religion (the Bible and Quran), politics, economics, sports, news, among others.

[1] https://www.press.et/
[2] https://www.fanabc.com/
[3] https://waltainfo.com/
[4] http://catalog.elra.info/en-us/repository/browse/ELRA-W0074/
[5] https://github.com/asmelashteka/HornMT
[6] https://github.com/adtsegaye/Amharic-English-Machine-Translation-Corpus
[7] https://github.com/admasethiopia/parallel-text

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

Authors	Trans. direction	# of sentence	Methods used	BLEU score
Biadgligne and Smaïli (2021)	En→Am	225,304	Statistical machine translation	26.47
			Neural machine translation	32.44
Gezmu et al (2021)	Am→En	45,364	Phrase-based statistical MT	20.2
			Neural Machine Translation	26.6
Abate et al (2018)	Am→En	40,726	Statistical machine translation	22.68
	En→Am		Statistical machine translation	13.31
Teshome et al (2015)	En→Am	18,432	Phoneme-based statistical MT	37.5
Teshome and Besacier (2012)	En→Am	18,432	Phrase-based statistical MT	35.32
Ashengo et al (2021)	En→Am	8,603	Context-based MT (CBMT) with RNN	11.34
Ambaye and Yared (2000)	En→Am	37,970	Statistical machine translation	18.74
Hadgu et al (2020)	Am→En	977	Google translate, Yandex translate	23.2, 4.8
	En→Am	1915	Google translate, Yandex translate	9.6, 1.3
Belay et al (2022)	Am→En	888,837	Transformer, M2M100 pre-trained model	16.26, 37.79
	En→Am	888,837	Transformer, M2M100 pre-trained model	13.06, 32.74

Note: **Trans. direction** column is the translation direction, "Am" is Amharic language and "En" is its English translation.

Table 2 Available Amharic-English parallel data sources

Data source	# of sentence pairs	Accessible?
Am-En ELRA-W0074[4]	13,347	yes
Biadgligne and Smaïli (2021)	225,304	yes
Horn MT[5]	2,030	yes
Am-En MT corpus[6]	53,312	yes
Gezmu et al (2021)	145,364	yes
Abate et al (2018)	40,726	yes
Lison and Tiedemann (2016)	562,141	yes
	60,884	no
Admasethiopia[7]	153	yes
Hadgu et al (2020)	2,914	yes
Newly curated (our data)	**33,955**	**yes**
Total	1,140,130	yes
Unique sentence pairs	888,837	yes

As we can see in Table 2, the total number of parallel sentences is around 1.1 million, while the unique parallel sentences are 888,837. This is due to duplication in the sources we used. This unique parallel sentence set is the largest to date.

The distribution of the Amharic-English dataset by year is shown in Figure 2. The duplicated year in the graph indicates the number of works done in that year. As it can be seen from the graph, in 2016, more Open Parallel Corpus (OPUS) data was collected (Lison and Tiedemann, 2016). In 2021, two basic MT works were done that focused on the Amharic-English parallel dataset (Biadgligne and Smaïli (2021); Gezmu et al (2021)). Until this data is compiled, our data is the first large-scale Amharic-English MT dataset.

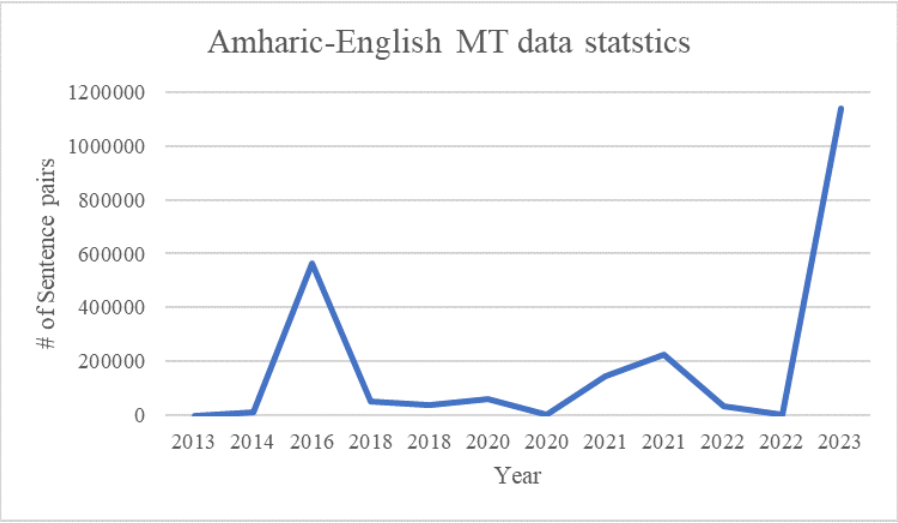

Fig. 2 Amharic-English MT data statistics by year

4.2 Data pre-processing

As our data was collected from different sources, we noticed a lot of textual irregularities. Some data was too noisy, so we eliminated it from our corpus. Some of the datasets are only available as raw data, no experiments have been conducted and reported on them yet such sets Am-En ELRA-W0074[8], Horn MT[9], Am-En MT corpus[10], and Admasethiopia[11]. So, these data need detailed pre-processing to become applicable for machine translation experiments.

We performed a series of pre-processing steps to canonize all tokens in Amharic and English sentences that were collected from various sources. The data preprocessing pipelines are shown in Figure 1. In this step, the following tasks were performed: data cleaning (removing URLs, hashtags, emojis, and duplicate sentences), abbreviation expansion for both Amharic and English, Amharic homophone character normalization, and English character lowercase.

For abbreviation expansion, we created a list of known abbreviations for both Amharic and English, then expanded them to their full writing form. Most English abbreviations were collected from available GitHub repositories[12]. In appendix D we included sample abbreviation lists we created for Amharic.

Segmentation of sentences essentially involves the disambiguation of end-of-sentence punctuation. For Amharic, we have used the available Python-based Amharic sentence segmentation module (pip install amseg) (Yimam et al, 2021; Belay et al, 2021).

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 meanings. The current trend in Amharic NLP research is, in most cases, to normalize these homophone characters into a single representation, which is called Amharic homophone character normalization; in some works, this trend is not applied. We released the dataset in both normalized and unnormalized forms. Amharic homophone characters are included in Appendix B2. The Amharic homophone character distributions in the dataset are shown in Appendix C3.

So, our corpus contains source sentences in Amharic and English. It is available in train, test, and validation split files that were experimented with in our previous work on neural machine translation (Belay et al, 2022). The data is split according to a 70:20:10 split strategy (training, testing, and validation, respectively). The datasets detail description are shown in Table 2. The total corpus consists of 15,673,082 and 19,820,336 tokens (words) for the Amharic and English languages, respectively.

[12] https://github.com/JRC1995/https://github.com/JRC1995/Machine-Translation-Transformers
[9] http://catalog.elra.info/en-us/repository/browse/ELRA-W0074/
[10] https://github.com/asmelashteka/HornMT
[11] https://github.com/adtsegaye/Amharic-English-Machine-Translation-Corpus
[12] https://github.com/admasethiopia/parallel-text

Table 3 Parallel sentences (segments) and distribution of tokens in each split dataset

			Amh Unnormalized[1]		Amh Normalized[2]	
Dataset	Sentences	En tokens	Tokens	Unique	Tokens	Unique
train set	774,848	14,267,853	11,276,110	571,391	11,280,417	553,658
test set	215,236	3,973,676	3,140,258	311,852	3,141,418	302,459
validation set	86,095	1,578,809	1,250,822	185,221	1,251,249	180,014
total	1,076,179	19,820,336	15,667,188	656,018	15,673,082	636,329

[1]Amh Un-Normalized is regular Amharic tokens without applying homophone normalization

[2]Amh Normalized is Amharic tokens after applying Amharic homophone normalization

5 Baseline NMT Experiments

This section describes NMT models used for baseline experiments along with their hyper-parameter configurations and the result of our baseline experiments on both unnormalized and normalized datasets.

5.1 NMT models

To investigate the usability of our dataset, we ran experiments in both datasets with recurrent neural networks (RNN) and Transformers with different hyper-parameter settings discussed in Section 5.2. We performed baseline experiments for bi-directional Amharic-English MT on both unnormalized and normalized datasets.

- **RNN**: is a class of artificial neural networks where connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes. We employed RNNs for the recurrent NMT model to build the internal representations of both the encoder and decoder in three different hyper-parameter configurations for our baseline experiments (Bahdanau et al, 2014). We implemented long short-term memory (LSTM) in recurrent layers.
- **Transformer**: is a new architecture that aims to solve tasks sequence-to-sequence while easily handling long-distance dependencies (Vaswani et al, 2017). 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. Similarly, as RNNs, we used Transformer in three different configurations for our baseline experiments.

5.2 Hyper-parameters

For the RNN model, we used RNN small, RNN base, and RNN large models based on the parameter settings shown in Table 4. Similarly, for the

Transformer model, we used the Transformer small, Transformer base, and Transformer large, with different parameter settings as shown in Table 4. For all the models, we used a learning rate of 0.0001, a batch size of 64, a total training example of 721622, a vocabulary size of 12086 with a maximum sentence length of 80, and ran for 250k steps per model. We used the Adam optimizer with a dropout of 0.1 for all experiments.

Table 4 RNN and Transformers hyper-parameters configuration

RNN			Transformer		
models	Hyper-param	value	models	Hyper-param	value
Small	num_layers	2	Small	num_heads	2
	hidden sizes	512		num_units	64
	# of parameters	29,384,040		inner_dim	64
				# of parameters	2,503,222
Base	num_layers	4	Base	num_heads	16
	encoder hidden size	256		num_units	512
	decoder hidden size	512		inner_dim	2048
	# of parameters	34,110,824		# of parameters	63,292,776
Large	num_layers	4	Large	num_layers	16
	encoder hidden size	512		num_units	1024
	decoder hidden size	1024		inner_dim	4096
	# of parameters	81,235,304		# of parameters	214,653,288

Note: Small, Base, and Large represent experimental configurations in both RNN and Transformer models

Table 4 shows the hyper-parameters used for RNN and Transformer models in different settings.

5.3 Experimental setting

We trained all models in Google Colab Pro + with OpenNMT (Klein et al, 2017) in the same environment as in Tonja et al (2022). We employed the BPE (Gage, 1994) subword for tokenization. We evaluated the performance of our baseline NMT models using Bi-lingual Evaluation Understudy (BLEU) score (Papineni et al, 2002), in terms of translation accuracy.

5.4 Results

Table 5 depicts the BLEU score results of two baseline models with three different parameter configurations. As depicted in Table 5 for RNN models, the performance of the RNN small is lower than the other two RNN models, and the RNN large outperformed the others. In the same way, Transformer Large outperformed the others in both datasets and translation directions for Amharic-English translation. This clearly shows that models with larger parameters can give better results than models trained with smaller parameters. From the results, we can also see that, from the two datasets used in this

experiment, the result of the normalized dataset is better than the result of the unnormalized dataset in all models. This means that the Amharic-English MT will work better if homophone normalization is used in Amharic sentences.

Table 5 Bi-directional Amharic-English NMT BLEU score results in Unnormalized and normalized dataset

	UnNormalized BLEU score[1]		Normalized BLEU score[2]	
Model name	amh-eng	eng-amh	amh-eng	eng-amh
RNN Small	12.14	9.84	12.38	10.20
RNN Medium	16.36	12.53	18.49	13.47
RNN Large	**23.70**	**13.21**	**26.42**	**14.06**
Transformer Small	14.78	10.79	16.26	12.06
Transformer Base	19.52	11.49	22.40	13.93
Transformer Large	**27.04**	**13.67**	**33.68**	**16.43**

[1]Amh Unnormalized is regular Amharic tokens without applying homophone normalization

[2]Amh Normalized is Amharic tokens after applying Amharic homophone normalization

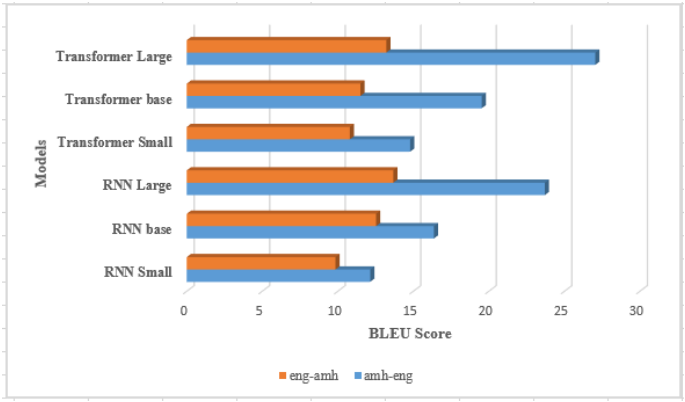

Fig. 3 Bi-directional Amharic-English NMT BLEU score results in unnormalized dataset

Figure 5 shows the comparison between two outperforming models from our baseline experiments. As we can see from the result, RNN and Transformer with large hyper-parameter configurations outperformed other models with small and medium hyper-parameter configurations. With a large number of hyper-parameters, Transformer performed better than RNN-large in both directions on both normalized and unnormalized datasets.

6 Conclusion

In this paper, we presented the first large and publicly available Amharic-English parallel dataset for machine translation. We collected, pre-processed,

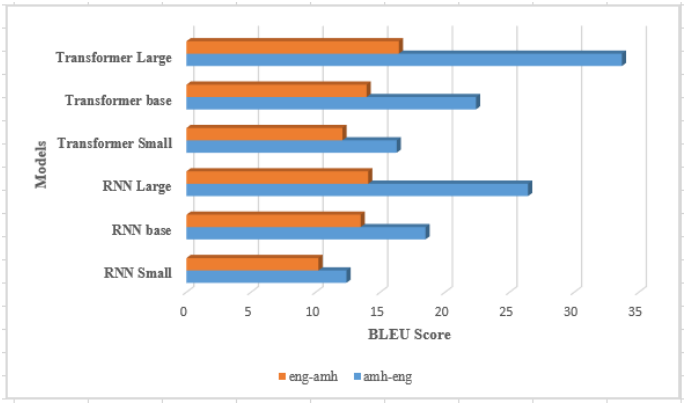

Fig. 4 Bi-directional Amharic-English NMT BLEU score result in normalized dataset

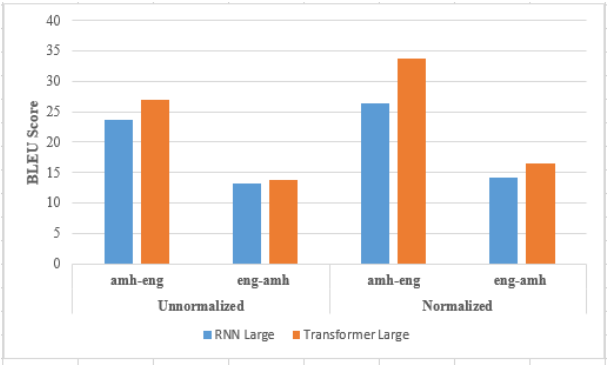

Fig. 5 Comparison of RNN Large and Transformer Large models in both datasets

segmented, and aligned Amharic-English parallel sentences from various sources. We conducted different baseline experiments to evaluate the usability of the collected corpus in bi-directional Amharic-English translation. We also showed the effect of Amharic homophone characters in the Amharic-English translation.

Acknowledgments. The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America Ph.D. Award.

Declarations

- Funding: This research received partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of Instituto Politécnico Nacional, Mexico.

- Conflict of interest/Competing interests: The authors declare no conflict of interest.
- Ethics approval: Not applicable.
- Consent to participate : Not applicable.
- Consent for publication : Not applicable.
- Availability of data and materials This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
- Code availability : will be available after acceptance of the manuscript
- Authors' contributions : conceptualization: A.L.T. and T.D.B., Methodology: A.L.T., T.D.B., O.K., S.M.Y., A.A.A, A.G. and G.S., Resources: A.L.T., T.D.B and O.K., Data curation: A.L.T., and T.D.B., Writing original draft: A.L.T. and T.D.B., Writing review and editing: A.L.T., T.D.B, O.K., A.A.A, S.M.Y, A.G and G.S., Supervision: O.K. and A.G.; Project administration: S.M.Y, O.K. and G.S., All authors reviewed the manuscript.

Appendix A Amharic Alphabets

Appendix B Amharic homophone characters

Appendix C Amharic homophone character distributions in the dataset

Appendix D Amharic Abbreviation samples

References

Abate ST, Melese M, Tachbelie MY, et al (2018) Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 3102–3111, URL https://aclanthology.org/C18-1262

Ambaye T, Yared M (2000) English to Amharic machine translation using statistical machine translation. Master's thesis

Ashengo YA, Aga RT, Abebe SL (2021) Context based machine translation with recurrent neural network for English–Amharic translation. Machine Translation 35(1):19–36. https://doi.org/10.1007/s10590-021-09262-4

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:14090473

Belay TD, Ayele AA, Gelaye G, et al (2021) Impacts of homophone normalization on semantic models for amharic. In: 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp 101–106, https://doi.org/10.1109/ICT4DA53266.2021.9672229

Belay TD, Tonja AL, Kolesnikova O, et al (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), IEEE, pp 84–89

Biadgligne Y, Smaïli K (2021) Parallel corpora preparation for English-Amharic machine translation. In: International Work-Conference on Artificial Neural Networks, Springer, Cham, pp 443–455, https://doi.org/10.1007/978-3-030-85030-2_37

Cho K, Van Merriënboer B, Bahdanau D, et al (2014) On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259

Eberhard DM, Simons GF, Fennig CD (2022) Ethnologue: Languages of the world (2022). URL URL:https://www.ethnologue.com/

Forcada ML, Ginestí-Rosell M, Nordfalk J, et al (2011) Apertium: a free/open-source platform for rule-based machine translation. Machine translation 25(2):127–144. https://doi.org/10.1007/s10590-011-9090-0

Gage P (1994) A new algorithm for data compression. C Users Journal 12(2):23–38

Gasser M (2011) Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In: Conference on Human Language Technology for Development, Alexandria, Egypt

Gehring J, Auli M, Grangier D, et al (2016) A convolutional encoder model for neural machine translation. arXiv preprint arXiv:161102344

Gezmu AM, Nürnberger A, Bati TB (2021) Extended parallel corpus for Amharic-English machine translation. arXiv preprint arXiv:210403543

Hadgu AT, Beaudoin A, Aregawi A (2020) Machine translation evaluation dataset for amharic. https://doi.org/10.5281/zenodo.3734260, URL https://doi.org/10.5281/zenodo.3734260

Hirschberg J, Manning CD (2015) Advances in natural language processing. Science 349(6245):261–266

Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing. Association for Computational Linguistics, Seattle, Washington, USA, pp 1700–1709

Klein G, Kim Y, Deng Y, et al (2017) Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:170102810

Koehn P, Hoang H, Birch A, 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. Association for Computational Linguistics, pp 177–180

Lison P, Tiedemann J (2016) OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp 923–929, URL https://aclanthology.org/L16-1147

Papineni K, Roukos S, Ward T, et al (2002) 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

Teshome MG, Besacier L (2012) Preliminary experiments on English-Amharic statistical machine translation. In: Spoken Language Technologies for Under-Resourced Languages, Cape Town, South Africa, pp 36–41, URL https://www.isca-speech.org/archive/

Teshome MG, Besacier L, Taye G, et al (2015) Phoneme-based English-Amharic statistical machine translation. In: AFRICON 2015, IEEE, Addis Ababa, Ethiopia, pp 1–5, https://doi.org/10.1109/AFRCON.2015.7331921

Tonja AL, Woldeyohannis MM, Yigezu MG (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), IEEE, pp 71–76

Tonja AL, Kolesnikova O, Arif M, et al (2022) Improving neural machine translation for low resource languages using mixed training: The case of ethiopian

languages. In: Mexican International Conference on Artificial Intelligence, Springer, pp 30–40

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30:5998–6008

Yigezu MG, Woldeyohannis MM, Tonja AL (2021) 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, pp 89–94

Yimam SM, Ayele AA, Venkatesh G, et al (2021) Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. Future Internet 13(11). https://doi.org/10.3390/fi13110275, URL https://www.mdpi.com/1999-5903/13/11/275

	Ge'ez ä	*Ka'eb* u	*Salis* ï	*Rab'e* a	*Hamis* é	*Sadis* i	*Sab'e* o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
a	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ạ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ts	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Fig. A1 Amharic Alphabets

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 (ha)	ኁ (hu)	ኂ (hi)	ኃ (hā)	ኄ (hé)	ኅ (he/h)	ኆ (ho)
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. B2 Amharic homophone characters

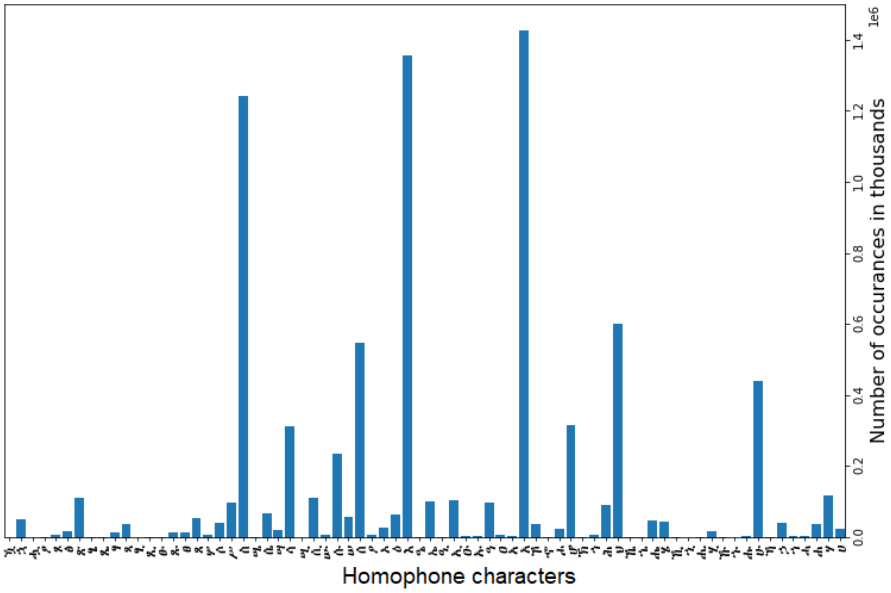

Fig. C3 Amharic homophone characters frequency

Abbreviations	Abbreviations expansion
ሆ/ል	ሆስፒታል
መ/ቤት	መሥሪያ ቤት
ሚ/ሩ	ሚኒስትሩ
ም/	ምክትል
ም/ቤት	ምክር ቤት
ሠ/ፌዴሬሽን	ሠራተኛ ፌዴሬሽን
ቤ/መ	ቤተ መንግስት
ቤ/ክ	ቤተ ክርስትያን
ተ/	ተከለ
ኃ/	ኃይለ
አ/አ	አዲስ አባባ
ኮ/ል	ኮለኔል
ወ/	ወልደ
ወ/ሪት	ወይዘሪት
ወ/ሮ	ወይዘሮ
ዓ/ም	ዓመተ ምህረት
ዓ/ዓ	ዓመተ ዓለም
ዶ/ር	ዶክተር
ጄ/ል	ጄኔራል
ጠ/ሚ	ጠቅላይ ሚኒስትር
ጠ/ሚ/ቢሮ	ጠቅላይ ሚኒስትር ቢሮ
ጠ/ሚኒስትር	ጠቅላይ ሚኒስትር
ጽ/ቤት	ጽህፈት ቤት
ፍ/ቤት	ፍርድ ቤት
ፕ/ር	ፕሮፌሰር
ፕ/ት	ፕሬዚዳንት
ት/ቤት	ትምህርት ቤት

Fig. D4 Amharic Abbreviations list