AGE: Amharic, Ge'ez and English Parallel Dataset

Henok Biadglign Ademtew
Ethiopian AI Institute
henokb2124@gmail.com

Mikiyas Girma Birbo
Maharishi International University
mbirbo@miu.edu

Abstract

African languages are not well-represented in Natural Language Processing (NLP). The main reason is a lack of resources for training models. Low-resource languages, such as Amharic and Ge'ez, cannot benefit from modern NLP methods because of the lack of high-quality datasets. This paper presents AGE, an open-source tripartite alignment of Amharic, Ge'ez, and English parallel dataset. Additionally, we introduced a novel, 1,000 Ge'ez-centered sentences sourced from areas such as news and novels. Furthermore, we developed a model from a multilingual pre-trained language model, which brings 12.29 and 30.66 for English-Ge'ez and Ge'ez to English, respectively, and 9.39 and 12.29 for Amharic-Ge'ez and Ge'ez-Amharic respectively.

1 Introduction

Language is fundamental to communication, with machine translation (MT) facilitating human-machine and human-human interactions (Abate et al., 2019). Data availability distinguishes high-resource from low-resource languages (Ranathunga et al., 2021). To date, there is no publicly available MT system for Ge'ez language and it's not represented in commercial MT systems such as Lesan[1], Google Translate[2], Microsoft Translator[3], and Yandex Translate[4]. It is also not included in large-scale pre-trained multilingual models like NLLB(Team et al., 2022), MT5(Xue et al., 2021), ByT5(elalliance, 2022), and M2M-100(Fan et al., 2020). This dataset aims to bridge historical linguistic heritage and modern technology, advancing MT capabilities and linguistic studies while contributing to the preservation of low-resource languages.

[1]https://lesan.ai
[2]http://translate.google.com/
[3]https://www.microsoft.com/en-us/translator/
[4]https://translate.yandex.com

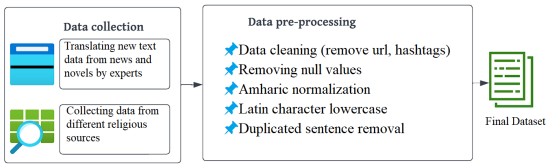

Figure 1: Data collection and pre-processing pipelines.

2 Related work

One of the major challenges in developing MT models for Ge'ez is the lack of public data. There were attempts to compile parallel corpora for Ge'ez to English and Ge'ez to Amharic MT tasks, but the development was unsatisfactory.

A recurring issue noted in these experiments is the absence of data sharing with the public domain. As shown in Table 1, there is a lack of open-sourcing data and models, a significant obstacle to the representation of Ge'ez in NLP.

3 Creation of the dataset

We introduce our newly Ge'ez-centered parallel dataset; **AGE** — **A**mharic, **G**e'ez , **E**nglish for machine translation.

We created a novel parallel dataset with 1,000 sentence pairs, later expanding it to 17,585 Amharic-Ge'ez and 18,676 Ge'ez-English pairs sourced from The Open Siddur Project, YouVersion, Ethiopic Bible, and Awde Mehret. Due to significant textual inconsistencies, we removed excessively disordered portions as shown in Figure 1. The dataset development involved collecting tripartite parallel sentences and translating some to Ge'ez using translators and evaluators. We developed an in-house tool to streamline this process and standardized tokens by cleaning the data, normalizing Amharic homophones, and converting English characters to lowercase.

Language	Sentences	Dataset	Model	Technique
Amharic, Ge'ez (Mulugeta, 2015)	12, 840	✗	✗	SMT
Amharic, Ge'ez (Kassa, 2018)	13,833	✗	✗	SMT
Amharic, Ge'ez (Abel, 2018)	976	✗	✗	SMT
Ge'ez, English (Abate et al., 2019)	11,663	✓	✗	SMT
Ge'ez, English (Getachew and Yayeh, 2023)	16,569	✗	✗	NMT
Amharic, Ge'ez (Tegenaw et al., 2023)	33,004	✗	✗	NMT
Amahric, Ge'ez (Wassie, 2023)	4,000	✗	✗	MNMT

Table 1: Summary of related works for Ge'ez, Sentences shows the number of sentences used during the experiment. Dataset and Model show the availability of datasets and models in publicly accessible repositories, and Technique shows the method used to build models.

Language pair	BLEU
Amharic-Ge'ez	9.39
Ge'ez-Amharic	12.29
English-Ge'ez	12.87
Ge'ez-English	30.66

Table 2: Baseline results of NLLB-200 600M

4 Baseline Experiments

Prior research predominantly used SMT and few employed NMT with transformers. We extended these studies using the NLLB-200(Team et al., 2022), a 54B parameter Mixture-of-Experts(MoE) model, but due to computational constraints, we utilized the NLLB-200 600M parameter variant. This model, fine-tuned on our dataset split into TRAIN (80%), DEV (10%), and TEST (10%), was trained using HuggingFace Transformer tool(Wolf et al., 2020) with specific parameters on Google Colab Pro. The training parameters included a learning rate of 5e-5, a batch size of 4 per device, a maximum source length, a maximum target length of 128, and a beam size of 10.

5 Results and Discussion

We adopted the NLLB-200 600M model to evaluate its performance in translating Ge'ez, as shown in Table 2, achieving BLEU scores of 9.39 and 12.26 for Amharic-Ge'ez and 30.35 and 30.66 for Ge'ez-English. Higher scores for English translations highlight the advantages of richer linguistic resources and extensive pre-training on English data (Team et al., 2022). Challenges in translating morphologically rich languages like Ge'ez were noted (Tran et al., 2014). This study presents the first ready-to-use Amharic, Ge'ez, English tripartite dataset, which will be made open source for further research. Future work will expand the dataset's quantity and diversity, incorporating more Ge'ez data sources.

References

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.

Biruk Abel. 2018. Geez to amharic machine translation.

elalliance. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Preprint*, arXiv:2105.13626.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Sefineh Getachew and Yirga Yayeh. 2023. Gex'ez-english bi-directional neural machine translation using transformer. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 160–164.

Tadesse Kassa. 2018. Morpheme-based bi-directional ge'ez -amharic machine translation.

Dawit Mulugeta. 2015. Geez to amharic automatic machine translation: A statistical approach.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ermias Tegenaw, Kris Calpotura, and Ashebir Dereje. 2023. Ge'ez to amharic translation with neural network-based technique.

Ke M. Tran, Arianna Bisazza, and Christof Monz. 2014. Word translation prediction for morphologically rich languages with bilingual neural networks. In *Conference on Empirical Methods in Natural Language Processing*.

Aman Kassahun Wassie. 2023. Machine translation for ge'ez language. *Preprint*, arXiv:2311.14530.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.