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

Tadesse Destaw Belay¹, Atnafu Lambebo Tonja², Olga Kolesnikova², Seid Muhie Yimam ³, Abinew Ali Ayele⁴, Silesh Bogale Haile⁵, Grigori Sidorov², and Alexander Gelbukh²

¹College of Informatics, Wollo University, Kombolcha, Ethiopia
tadesseit@gmail.com
²Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexicoalamedot2022@cic.ipn.mx, kolesolga@gmail.com, (lastname)cic.ipn.mx
³Dept. of Informatics, Universität Hamburg, Hamburg, Germanyseid.muhie.yimam@uni-hamburg.de
⁴Faculty of Computing, Bahir Dar University, Bahir dar, Ethiopiaabinewaliayele@gmail.com
⁵Dept. of Computer Science, Assosa University, Assosa, Ethiopia
sileshibogale123@gmail.com

Abstract

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

1 Introduction

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

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

The focus of MT research for Amharic¹ has been on rule-based and SMT methods. In this work, we used the transformer model as a baseline translation system to explore the applicability of the Facebook multilingual pre-trained model (Fan et al., 2020).

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

2 Methodology

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

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

¹Amharic is the official language of the Federal Democratic Republic of Ethiopia (FDRE) and for many regional states in the country.
parallel sentences are 888k. This is due to duplication in the sources we used. This unique parallel sentence is the largest to date. In addition to the available datasets in Table 2, we have contributed to the MT research field by creating a new parallel corpus with 33,955 sentence pairs extracted from such news platforms as Ethiopian Press Agency, Fana Broadcasting Corporate, and Walta Information Center. As the data we used is from different sources, it includes various domains such as religious, politics, economics, sports, news, among others.

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

Table 2: Amharic-English parallel data sources

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

*Not available freely

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

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

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

3 Experimental Setup and Results

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

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

<table>
<thead>
<tr>
<th>Models &amp; Trans. direction</th>
<th>Regular</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer Am→En</td>
<td>14.78</td>
<td>16.26</td>
</tr>
<tr>
<td>Transformer En→Am</td>
<td>10.79</td>
<td>13.06</td>
</tr>
<tr>
<td>M2M100 48M Am→En</td>
<td>34.12</td>
<td><strong>37.79</strong></td>
</tr>
<tr>
<td>M2M100 48M En→Am</td>
<td>29.65</td>
<td><strong>32.74</strong></td>
</tr>
</tbody>
</table>

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

4 Conclusion and Future Work

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

References


\(^5\)https://github.com/uhh-lt/ethiopicmodels


