Parallel Corpora for bi-Directional Statistical Machine Translation for Seven Ethiopian Language Pairs

Solomon Teferra Abate
Addis Ababa University,
Addis Ababa, Ethiopia
solomon.teferra@aau.edu.et

Michael Melese Woldeyohannis
Addis Ababa University,
Addis Ababa, Ethiopia
michael.melese@aau.edu.et

Martha Yifiru Tachbelie
Addis Ababa University,
Addis Ababa, Ethiopia
martha.yifiru@aau.edu.et

Million Meshesha
Addis Ababa University,
Addis Ababa, Ethiopia
million.meshesha@aau.edu.et

Solomon Atinafu
Addis Ababa University,
Addis Ababa, Ethiopia
solomon.atnafu@aau.edu.et

Wondwossen Mulugeta
Addis Ababa University,
Addis Ababa, Ethiopia
wondwossen.mulugeta@aau.edu.et

Yaregal Assabie
Addis Ababa University,
Addis Ababa, Ethiopia
yaregal.assabie@aau.edu.et

Hafte Abera
Addis Ababa University,
Addis Ababa, Ethiopia
hafte.abera@aau.edu.et

Biniyam Ephrem
Addis Ababa University,
Addis Ababa, Ethiopia
binyam.ephrem@aau.edu.et

Tewodros Abebe
Addis Ababa University,
Addis Ababa, Ethiopia
tewodros.abebe@aau.edu.et

Wondimagegnhue Tsegaye
Bahir Dar University,
Bahir Dar, Ethiopia
wendeal@gmail.com

Amanuel Lemma
Aksum University,
Axum, Ethiopia
amanu.infosys@gmail.com

Tsegaye Andargie
Wolkite University,
Wolkite, Ethiopia
adtsegaye@gmail.com

Seifedin Shifaw
Wolkite University,
Wolkite, Ethiopia
seifedin28@gmail.com

Abstract

In this paper, we describe the development of parallel corpora for Ethiopian Languages: Amharic, Tigrigna, Afan-Oromo, Wolaytta and Ge'ez. To check the usability of all the corpora we conducted baseline bi-directional statistical machine translation (SMT) experiments for seven language pairs. The performance of the bi-directional SMT systems shows that all the corpora can be used for further investigations. We have also shown that the morphological complexity of the Ethio-Semitic languages has a negative impact on the performance of the SMT especially when they are target languages. Based on the results we obtained, we are currently working towards handling the morphological complexities to improve the performance of statistical machine translation among the Ethiopian languages.

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 83–90
Santa Fe, New Mexico, USA, August 20, 2018.

1 Introduction

The advancement of technology and the rise of the internet as a means of communication led to an ever-increasing demand for Natural Language Processing (NLP) Applications. NLP applications are useful in facilitating human-human communication via computing systems. One of the NLP applications which facilitate human-human communication is Machine Translation. Machine Translation (MT) refers to a process by which computer software is used to translate a text from one language to another (Koehn, 2009). In the presence of high volume digital text, the ideal aim of machine translation systems is to produce the best possible translation with minimal human intervention (Hutchins, 2005).

The translation of natural language by machine becomes a reality, for technologically favored languages, in the late 20th century although it is dreamt since the seventieth century (Hutchins, 1995). Various approaches to MT have been and are being used in the research community. These approaches are broadly classified into rule based and corpus-based MT (Koehn, 2009). The rule-based machine translation demands various kinds of linguistic re- sources such as morphological analyzer and synthesizer, syntactic parsers, semantic analyzers and so on. On the other hand, corpus-based approaches (as the name implies) require parallel and monolingual corpora. Since corpus-based approaches do not require deep linguistic analysis of the source and target languages, it is the preferred approach for under-resourced languages of the world, including Ethiopian languages.

1.1 Machine Translation for Ethiopian Languages

Research in the development of MT has been conducted for technologically favored and economically as well as politically important languages of the world since the 17th century. As a result, notable progress towards the development and use of MT systems has been made for these languages. However, research in the area of MT for Ethiopian languages, which are under-resourced as well as economically and technologically disadvantaged, has started very recently. Most of the researches on MT for Ethiopian languages are conducted by graduate students (Tariku, 2004; Sisay, 2009; Eleni, 2013; Jabesa, 2013; Akubazgi, 2017), including two PhD works: one that tried to integrate Amharic into a unification-based machine translation system (Sisay, 2004) and the other that investigated English-Amharic Statistical Machine translation (Mulu, 2017). Beside this, Michael and Million (Michael and Million, 2017) experimented a bi-directional Amharic-Tigrigna SMT system using word and morpheme as a unit.

Due to unavailability of linguistic resources and since the most widely used MT approach is statistical, most of these researches have been conducted using statistical machine translation (SMT), which requires large bilingual and monolingual corpora. However, as there were no such corpora for SMT experiments for Ethiopian languages, the researchers had to prepare their own small size corpora for their experiment. This in turn, affects the results that they obtain.

In addition, since there are no standard corpora for conducting replicable and consistent experiments for performance evaluation, it is difficult to know the progress made in the area for local languages. Moreover, since the researchers had to spend their time on corpora preparation, they usually have limited time for experimentation, exploration and development of MT systems.

1.2 Motivation of this Paper

African languages, which contribute around 30% (2139) of the world language highly suffer from the lack of sufficient language resources (Simons and Fennig, 2017). This is true for Ethiopian languages. On the other hand, Ethiopia being a multilingual and multi-ethnic country, its constitution decrees that each citizen has the right to speak, write and develop in his/her own language. However, there is still a need to share information among citizens who speak different languages. For example, Amharic is the regional language of the Amhara and Southern Nations and Nationalities regions, Afan-Oromo is that of the Oromia region while the Tigray region uses Tigrigna. All these regions produce a lot of information that need to be shared among the other regions of the nation. There is, therefore, a lot of translation demands among the different language communities of the federal government of Ethiopia.

In order to enable the citizens of the country to use the documents and the information produced in other Ethiopian languages, the documents need to be translated to the languages they understand most. Since manual translation is expensive, a promising alternative is the use of machine translation, particularly SMT as Ethiopian languages suffer from lack of basic linguistic resources such as morphological analyser, syntactic analyser, morphological synthesizer, etc. The major and basic

resource required for SMT is parallel corpora, which are not available for Ethiopian languages. The collection and preparation of parallel corpora for Ethiopian languages is, therefore, an important endeavour to facilitate future MT research and development.

We have, therefore, collected and prepared parallel corpora for seven Ethiopian Language pairs taking representatives from the Semitic, Cushitic and Omotic language families. We have considered Amharic, Tigrigna and Ge'ez from the Semitic, Afan-Oromo from the Cushitic and Wolaytta from Omotic language families. This paper, therefore, describes the parallel corpora we collected and prepared for these Ethiopian languages and the SMT experiments conducted using the corpora as a way of verifying their usability.

2 Nature of the Language Pairs

The language pairs in the corpora belong to Semitic (Ge'ez, Amharic and Tigrigna), Cushitic (Afan-Oromo) and Omotic (Wolaytta) language families. Except Ge'ez, these languages have native speakers. Ge'ez serves as a liturgical language of Ethiopian Orthodox Church. It is thought as a second language in the traditional schools of the church and given as a course in different Universities. There is a rich body of literature in Ge'ez, including philosophical, medical and astrological writings. Because of this, there is a big initiative in translating the documents written in Ge'ez to other widely used languages. On the other hand, Amharic is spoken by more than 27 million people which makes it the second most spoken Semitic language in the world. Tigrigna is spoken by 9 million people. Afan-Oromo and Wolaytta are spoken by more than 34 million and 2 million speakers, respectively (Simons and Fennig, 2017).

The writing systems of these language pairs are Ge'ez or Ethiopic script and Latin alphabet. Ge'ez, Amharic and Tigrigna are written in Ge'ez script whereas both Afan-Oromo and Wolaytta are written in Latin alphabet. It is believed that the earliest known writing in the Ge'ez script date back to the 5th century BC. The Ge'ez script is syllabary in which each character represents a consonant and a vowel. Each character gets its basic shape from the consonant of the syllable, and the vowel is represented through systematic modifications of the basic shape. The script is also used to write other languages like Argobba, Harari, Gurage, etc.

The language pairs have got different functions in the country. Amharic for instance is the working language of the Federal Government of Ethiopia. It also serves as regional working language of some other regional states. It facilitates inter-regional communication. Tigrigna and Afan-Oromo are working languages in Tigray and Oromia regional administrations, respectively. Apart from this, they serve as medium of instructions in primary and secondary schools. These languages are also used widely in the electronic media like news, blogs and social media. Some of the governmental websites are available in Amharic, Tigrigna and Afan-Oromo. Currently, Google offers a searching capability using these Ethiopian languages. Further, Google also included Amharic in its translation service recently.

2.1 Morphological Features

Like other Semitic languages, Ge'ez (Dillmann and Bezold, 1907), Amharic (Leslau, 2000; Anbessa and Hudson, 2007) and Tigrigna (Mason, 1996; Yohannes, 2002), make use of the root and pattern system. In these languages, a root (which is called a radical) is a set of consonants which bears 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 results in derived words. Such derivational process makes these languages morphologically complex.

In addition to the morphological information, some syntactic information are also expressed at word level. Furthermore, an orthographic word may attach some syntactic words like prepositions, conjunctions, negation, etc. which create various word forms (Gasser, 2010; Gasser, 2011). In these languages, nominals are inflected for number, gender, definiteness and case whereas verbs are inflected for person, number, gender, tense, aspect, and mood (Griefenow-Mewis, 2001).

Essentially, unlike the Semitic languages which allow prefixing, Afan-Oromo allows suffixing. Most functional words like postpositions are also suffixed. However, there are some prepositions written as a separate word.

Wolaytta like Afan-Oromo is a suffixing language in which words can be generated from root words recursively by adding suffixes only. Wolaytta nouns are inflected for number, gender and case whereas verbs are inflected for person, number, gender, aspect and mood (Wakasa, 2008).

2.2 Syntactic Features

Ethiopian languages that are under our consideration follow Subject-Object-Verb (SOV) word-order except Ge'ez which allows the verb to come first. In Ge'ez, the basic word-order is Verb-Subject-Object (VSO).

3 Challenges of SMT

Statistical Machine Translation is greatly impacted by the linguistic features of the target languages. The challenges range from the writing system to that of word ordering and morphological complexity.

3.1 Writing System

The Ge'ez writing system, which is used by Amharic, Tigrigna and Ge'ez languages, uses different characters in words that convey the same meaning, especially in Amharic. For example, peace can be written as: ሰላም or ሠላም. Such character variations affect probability values that have direct impact on the performance of SMTs.

3.2 Word Ordering

Most of the languages under consideration have same word order. With this respect, Amharic, Afan-Oromo, Tigrigna and Wolaytta have SOV, while only Ge'ez has VSO. This might challenge machine translation system where Ge'ez is in the pair. Another challenge is the existence of flexibility in word order. For instance, even though Afan-Oromo follows SOV word order, nouns can be changed based on their role in a sentence which makes the word order to be flexible. Such flexibility will pose a challenge for translation from a source to Afan-Oromo.

3.3 Morphological Complexity

While word alignment could be done automatically or with supervision, morphological agreement between words in the source and target are crucial. For instance, Amharic and Geez have subject agreement, object agreement and genitive (possessive) agreement. Each of which is expressed as bound morphemes. In Amharic, for the word ገድልክ/you killed/ the subject "you" is represented by the suffix "+ክ" while the same subject is represented as "+" in the Geez ቀተልከ /you killed/). Most of the morphemes in the considered Ethiopian languages are bound ones.

4 Parallel Corpora Preparation

The development of machine translation more often uses statistical approach because it requires very limited computational linguistic resources compared to the rule based approach. Nevertheless, the statistical approach relies to a great extent on parallel corpora of the source and target languages.

The research team has applied different techniques to collect parallel corpora for the selected Ethiopian language pairs. The domain of the collected data is only religious for which we have data for all the considered language pairs. It includes Holy Bible and different documents written in spiritual theme and collected from Jehovah's Witnesses (JW[1]), Ethiopicbible[2], Ebible[3] and Geez experience[4] which are freely accessible websites.

A simple web crawler was used to extract parallel text from the websites. Python libraries such as requests, and BeautifulSoup were used to analyse the structure of the websites, extract texts

[1] available at https://www.jw.org
[2] available at https://www.ethiopicbible.com
[3] available at http://ebible.org
[4] available at https://www.geezexperience.com

and combine into a single text file. To collect the bible data, we have generated the structure of the URL so that it shows the book names, chapters and verses of the Bible in each language.

For the "daily text" which is published at JW.org, we tried to use the date information to generate URL for each language. Finally, we extracted the data based on the generated URL information and merged to a single UTF-8 text file for each language.

4.1 Pre-processing

Data preprocessing is an important and basic step in preparing bilingual and multilingual parallel corpora. Since the collected parallel data have different formats and characteristics, it is very difficult and time-consuming to prepare usable parallel corpora manually because it needs to analyse the structure of the collected raw data by applying different linguistic methods. We have, therefore, applied different automatic methods of text pre-processing that requires minimal human interference. As part of the pre-processing unnecessary links numbers, symbols and foreign texts in each language have been removed. During pre-processing the following tasks have been performed: character normalization, sentence tokenization and sentence level alignment.

4.1.1 Character Normalization

As it is indicated in Section 3.1, there are characters in Amharic that have similar roles and are redundant. For example the character (ሀ) can be written as (ሐ, ሓ, ኀ, ኻ and ሃ). Though they used to possess semantic differences in the traditional writings, currently these characters are mostly used interchangeably. To avoid words with same meaning from being taken as different words due to these variations we have replaced a set of characters with similar function into a single most frequently used character.

4.1.2 Sentence Tokenization and Alignment

Lines that contain multiple sentences in both source and target languages are tokenized. The team has set two criteria to check whether the aligned sentences are correct or not. The first criterion is counting and matching the number of sentences in the source language and the target language. In the parallel corpora of the language pairs in which Ge'ez is the target, the source language contains multiple verses in a single line. While on the Ge'ez side, each line contains a single verse. In such cases, we merged different verses of Ge'ez to produce the line that is aligned with that of the source language.

4.2 Corpus Size and Distribution of Words

The corpora have been analysed to see the relationship between languages in the language pairs. As it has been revealed in different literature, the Ethio-Semitic languages have more complex morphology than the other Ethiopian languages. Due to this difference, the same number of sentences in these language pairs is tokenized into significantly different number of tokens and word types. Table 1 clearly shows that the vocabulary of the languages in the Ethio-Semitic language family is much more than the vocabulary of the other two language families.

Sentences	Languages	Token	Type	Average sentence Length
34,349	Amharic	521,035	98,841	15
	Tigrigna	546,570	87,649	15
11,546	Amharic	148,084	38,097	12
	Ge'ez	158,003	33,386	13
11,457	Amharic	163,816	37,283	14
	Afan-Oromo	214,335	24,005	18
10,987	Tigrigna	162,508	32,953	14
	Afan-Oromo	206,844	23,536	18
9,400	Amharic	119,262	32,780	12
	Wolaytta	137,869	25,331	14
	Afan-Oromo	46,340	8,118	15

Sentences	Languages	Token	Type	Average sentence Length
2,923	Wolaytta	33,828	8,786	11
2,504	Tigrigna	34,780	9,864	13
	Wolaytta	29,458	7,989	11

Table 1: Sentence and Word Distribution of the Parallel Corpora

On the contrary the token of the non-Semitic languages is significantly higher than the tokens of the Ethio-Semitic languages. This is because syntactic words like preposition, conjunction, negation, etc are bound in the Ethio-Semitic language group. It is clear, therefore, that such differences between the languages in a language pair makes SMT difficult because it aggravates data sparsity and results into a weakly trained translation model. Although the size of the data we have is much less to draw conclusions, we could also see that the Ethio-Cushitic and Ethio-Omotic languages are morphologically more similar with each other than their similarity with the Ethio-Semitic languages.

We have also observed morphological differences among the Ethio-Semitic languages that is revealed by the difference in the number of token and word type in the same corpus we have for Amharic-Tigrigna and Amharic-Ge'ez language pairs. The data revealed that Amharic is the most morphologically complex language of the family.

5 SMT Experiments and Results

To check the usability of the collected parallel corpora for seven Ethiopian language pairs, we have conducted bi-directional SMT experiments.

5.1 Experimental Setup

To conduct SMT experiments, each parallel corpus has been divided into three sets: 80% for the training, 10% for tuning and 10% for test sets. Moses (Koehn, 2009) has been used along with Giza++ alignment tool (Och and Ney, 2003) for aligning words and phrases. SRILM toolkit (Stolcke, 2002) has been used to develop the language models using target language sentences from the training and tuning sets of parallel corpora. Bilingual Evaluation Under Study (BLEU) is used for automatic scoring.

5.2 Experimental Results

Table 2 presents the experimental results of bi-directional SMT systems developed for the seven Ethiopian language pairs. The Table shows the difference in the performance of the systems developed for the same language pair in different directions.

Sentences	Language pair	BLEU
34,349	Amharic - Tigrigna	21.22
	Tigrigna - Amharic	19.06
11,457	Amharic - Afan Oromo	17.79
	Afan Oromo - Amharic	13.11
10,987	Tigrigna - Afan Oromo	16.82
	Afan Oromo - Tigrigna	14.61
9,400	Amharic - Wolaytta	11.23
	Wolaytta - Amharic	7.17
11,546	Ge'ez - Amharic	7.31
	Amharic - Ge'ez	6.29
2,923	Wolaytta - Afan Oromo	4.73
	Afan Oromo - Wolaytta	2.73
2,504	Tigrigna - Wolaytta	2.2
	Wolaytta - Tigrigna	3.8

Table 2: Experimental Results

The performance of SMT systems decreases when Ethio-Semitic languages are on the target side. This confirms that Ethio-Semitic languages (when they are targets) are more challenging to SMT than the other language families. The only exception being Tigrigna and Wolaytta language pair, where the performance is high when Tigrigna is a target. This could be attributed to the small data we have for this language pair.

The results in the Table 2 also show the effect of data size on the performance of SMT systems. That means as the data increases, the performance also increases. In this view again we have an exceptionally lower BLUE score for the Amharic-Ge'ez language pair than the score we achieved for Amharic-Afan Oromo language pair although the data size used is almost equal. The performances of the Amharic-Wolaytta-Amharic translation systems are better than Amharic- Ge'ez-Amharic systems although the data size used in the former is less than the data size used in the latter. The results confirm that the morphological complexity of the languages severely affect SMT performance than the amount of data. From the difference in results achieved for the Amharic-Ge'ez (6.29) and Ge'ez-Amharic (7.31) language pairs, it is possible to understand that syntactic differences affect the performance of SMT more than the difference in their morphological features. We have seen from their number of word types that Amharic has more complex morphology than Ge'ez which, however, has flexible syntactic structure.

Despite the size of the data, the performance registered in translation towards the Ethio- Semitic languages has less BLEU score than the translations from them. This is because of the fact that, when the Ethio-Semitic languages, specially Amharic, are used as a target language, the translation from other languages with less morphological complexity as a source languages is challenged by one-to-many alignment. On the other hand, better performance is registered the other direction since the alignment is many-to-one. Beside this, the word-based language model favours the non-Semitic languages than Semitic ones due to the complexity of the morphology of the latter language family.

6 Conclusions and Future Work

This paper presents the attempt made in the preparation of usable parallel corpora for Ethiopian languages. The corpora have been collected from the web in the religious domain. Then, they are further pre-processed and normalized. We have now usable parallel corpora for seven Ethiopian language pairs. Using the corpora, bi-directional statistical machine translation experiments have been conducted. The results show that translation systems from Ethio-Semitic languages to either Omotic or Cushitic language families achieve better BLEU score than those in the other directions. That leads us to conclude that the Ethio-Semitic language family has the most complex morphology which greatly affects the performance of SMT.

Finding solutions that minimize the negative effect of morphological complexity of the Ethio-Semitic languages on the performance of SMT is therefore a future endeavour. We considered a previous work by (Mulu, 2017) who gained a significant improvement by the application of morphological segmentations to guide us in utilizing the use of morphemes instead of words as units for both the translation and the statistical language models. The most attractive solution to the problems of machine translation that is the trend of the time is the use of ANN modelling, which however, requires more data than what we have. So we will use our experience of corpus preparation and work towards the application of the state of the art technologies to develop usable machine translation systems for the Ethiopian languages.

As it is well known, domain is the most important factor on the performance of SMT. Thus, we are also working on the development and organization of parallel corpora for the Ethiopian languages in different domains.

Acknowledgement

We would like to acknowledge Addis Ababa University for the thematic research fund and NORHED (Linguistic Capacity Building-Tools) for sponsoring conference participation.

References

Gebremariam Akubazgi. 2017. *Amharic-Tigrigna machine translation using hybrid approach*. Master's thesis, Addis Ababa University.

Teferra Anbessa and Grover Hudson. 2007. *Essentials of Amharic*. Rüdiger Köppe, Verlag, Köln.

August Dillmann and Carl Bezold. 1907. *Ethiopic grammar, enlarged and improved by c. Bezold*. Translated by JA Crichton. London: Williams & Norgate.

Teshome Eleni. 2013. *Bidirectional English-Amharic machine translation: An experiment using con- strained corpus*. Master's thesis, Addis Ababa University.

Michael Gasser. 2010. A dependency grammar for Amharic. In *Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages*, Valletta, Malta. ftp://html.soic.indiana.edu/pub/gasser/sem_ws10.pdf

Michael Gasser. 2011. Hornmorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Conference on Human Language Technology for Development*, pages 94-99, Alexandria, Egypt.

Catherine Griefenow-Mewis. 2001. *A grammatical sketch of written Oromo*, volume 16. Rüdiger Köppe.

John Hutchins. 1995. Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445. Elsevier.

John Hutchins. 2005. The history of machine translation in a nutshell, http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

Daba Jabesa. 2013. *Bi-directional English-Afaan oromo machine translation using hybrid approach*. Master's thesis, Addis Ababa University.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Wolf Leslau. 2000. *Introductory grammar of Amharic*, volume 21. Otto Harrassowitz Verlag.

John S Mason. 1996. *Tigrinya grammar*. Red Sea Press (NJ).

Melese Michael and Meshesha Million. 2017. Experimenting Statistical Machine Translation for Ethiopic Semitic Languages : The case of Amharic-Tigrigna. In *LNICST. EAI International Conference on ICT for Development for Africa*, pages 140-149, Springer.

Gebreegziabher Teshome Mulu. 2017. *English-Amharic Statistical Machine Translation*. PhD Dissertation, IT Doctoral Program, Addis Ababa University, Addis Ababa, Ethiopia.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Gary F Simons and Charles D Fennig. 2017. *Ethnologue: Languages of the world*. SIL, Dallas, Texas.

Adugna Sisay. 2009. *English–Afan Oromo machine translation: An experiment using statistical Approach*. Master's thesis, Addis Ababa University.

Fissaha Sisay. 2004. *Adding Amharic to a Unification-based Machine Translation System: An Experiment*. Peter Lang.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, https://www.isca-speech.org/archive/icslp_2002/i02_0901.html

Tsegaye Tariku. 2004. *English-Tigrigna factored statistical machine translation*. Master's thesis, Addis Ababa University.

Motomichi Wakasa. 2008. *A descriptive study of the modern Wolaytta language*. Unpublished PhD thesis, University of Tokyo.

Tesfay Tewolde Yohannes. 2002. *A modern Grammar of Tigrinya*. Tipografia U. Detti.