Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation

Solomon Teferra Abate[1], Michael Melese Woldeyohannis[1], Martha Yifiru Tachbelie[1], Million Meshesha[1], Solomon Atinafu[1], Wondwossen Mulugeta[1], Yaregal Assabie[1], Hafte Abera[1], Biniyam Ephrem[1], Tewodros Abebe[1], Wondimagegnhue Tsegaye[2], Amanuel Lemma[3], Tsegaye Andargie[4], Seifedin Shifaw[4]

[1]Addis Ababa University, Addis Ababa, Ethiopia, [2]Bahir Dar University, Bahir Dar, Ethiopia
[3]Aksum University, Axum, Ethiopia, [4]Wolkite University, Wolkite, Ethiopia
{solomon.teferra, michael.melese, martha.yifiru, million.meshesha, solomon.atnafu, wondwossen.mulugeta, yaregal.assabie, hafte.abera, binyam.ephrem, tewodros.abebe}@aau.edu.et, {wendeal, amanu.infosys, adtsegaye, seifedin28}@gmail.com

Abstract

In this paper, we describe an attempt towards the development of parallel corpora for English and Ethiopian Languages, such as Amharic, Tigrigna, Afan-Oromo, Wolaytta and Ge'ez. The corpora are used for conducting a bi-directional statistical machine translation experiments. The BLEU scores of the bi-directional Statistical Machine Translation (SMT) systems show a promising result. The morphological richness of the Ethiopian languages has a great impact on the performance of SMT specially when the targets are Ethiopian languages. Now we are working towards an optimal alignment for a bi-directional English-Ethiopian languages SMT.

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-machine and human-human communications. One of the NLP applications which facilitates human-human communication is Machine Translation (MT). MT refers to a process by which computer software is used to translate a text or speech 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 20^{th} century although it is dreamt in seventieth century (Hutchins, 1995). Various approaches to MT have been and are being used in the research community, that can broadly classified as rule-based and corpus based (Koehn, 2009). The rule based machine translation demands various kinds of linguistic resources 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 17^{th} 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

Proceedings of the 27th International Conference on Computational Linguistics, pages 3102–3111
Santa Fe, New Mexico, USA, August 20-26, 2018.

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 these, Michael and Million (2017) attempted a bi-directional Amharic-Tigrigna SMT experiement using word and morpheme as translation units.

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

In addition, since there is no standard corpora for conducting replicable and consistent experiment for performance evaluation, it is difficult to know the progress made in the area for local languages. Moreover, since the researchers 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 languages, highly suffer from lack of sufficient language resources (Simons and Fennig, 2017). This is true for Ethiopian languages as well. On the other hand, Ethiopia being multilingual and multiethnic country, its constitution decrees that each citizen has the right to speak, write and develop in his/her own language. However, a lot of written documents, brochures, text books, magazines, advertisements and other information in the web are being produced in technological favored and economically important languages such as English.

In order to enable Ethiopians to use the documents and information produced in technologically favored languages, the documents need to be translated. 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 analyzer, syntactic analyzer, 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 endeavor to facilitate future MT research and development. Corpus acquisition for SMT is actually one of the recommendations of Saba and Sisay (2006).

We have, therefore, collected and prepared parallel corpora for English and Ethiopian Languages that fall under 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 an attempt that we have made to collect and prepare English-Ethiopian languages parallel corpora and the SMT experiments conducted using the corpora.

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. Presently, Ge'ez does not have native speakers. Ge'ez functions as a liturgical language of Ethiopian Orthodox Church. It is thought as a second language in traditional schools of churches and given as a course in different Universities. There is a rich body of literature in Ge'ez. It is not only literature but also philosophical, medical and astrological documents were written in Ge'ez. Because of this, there is a big initiative in translating those documents written in Ge'ez. On the other hand, Amharic is spoken by more than 27 million people which makes it the second most spoken Semitic language. 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 dated back to the 5^{th} century BC. The writing system is syllabry where each character represents a consonant and vowel. The basic features of the writing system is that each character gets its basic shape from the consonant of the syllable, and the vowel is represented through a systematic modifications of these basic shapes. The script is used to write other languages like Amharic, Tigrigna, Argoba, etc.

The Ethiopian languages considered in the project 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 as well. Tigrigna and Afan-Oromo are working language in Tigray and Oromiya regional administrations, respectively. Some of the governmental websites are available in Amharic, Tigrigna and Afan-Oromo. Apart from this, they serve as medium of instructions in primary and secondary schools. These languages are available in electronic media like news, blogs and social media except Ge'ez. Currently, Google offers a searching capability using Amharic, Tigrigna and Afan-Oromo. Further, Google also included Amharic in its translation service recently.

2.1 Morphological features

As in other Semitic language morphology, Ge'ez (Dillman, 1907), Amharic (Leslau, 2000; Teferra 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 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 patters together with affixes results in derived words. Such derivational process makes these language to be morphologically complex languages.

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 make word forms to be very varied (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.

As we may observe in the Semitic languages, nominals are inflected for number, gender, case and definiteness and verbs are inflected for person, number, gender, tense, aspect and mood (Griefenow-Mewis, 1995). Essentially, unlike the Semitic languages which allow prefixing, Afan-Oromo allows suffixing. Most functional words like pospositions 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). On the contrary, English language uses Subject-Verb-Object (SVO) word-order.

3 Challenges of SMT

Statistical machine translation is greatly affected by the linguistic features of the target languages. The challenges ranges from the writing system to that of word ordering and morphological complexity.

3.1 Writing System

Semitic language writing system are represented by a consonant vowel (CV) sequence and the basic shape of each character is determined by the consonant, which is modified for the vowel. These language script has inherited its writing system from Ge'ez (ግእዝ) /gə'əzə/ using a grapheme based writing system called fidel /fidälə/ which is written and read from left to right being the classical and ecclesiastical language of Ethiopia. Unlike the Semitic languages, The Cushitic (Afan-Oromo) and Omotic (Wolaytta) languages use a latin based writing system.

3.2 Word Ordering

Semitic languages like Amharic, Tigrigna and Ge'ez are morphologically rich where words are required to be further segmented or a single word from these languages would be aligned with as big as handful of words in languages like English. The languages under consideration have different word order. With this respect, Amharic, Afan-Oromo, Tigrigna and Wolaytta have SOV, Ge'ez has VSO and English has SVO topology. The different word orders used in these languages is major challenge for Multilingual machine translation system.

Another challenge is the existence of flexibility in word order. For instance, even though Afan-Oromo follows SOV word order format, nouns can be changed based on their role in a sentence which makes the word order to be flexible. Although the major word order of Ge'ez is VSO, it also follows free word order. Such flexibilities will pose another challenge for translation from source to Afan-Oromo and Ge'ez languages.

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 Ge'ez have subject agreement, object agreement and genitive (possessive) agreement. Each of which are expressed as bound morphemes which should be aligned or translated as independent words in English. In Amharic, for the word **ገደልክ** /you killed/ the English subject "you" is represented by the suffix "+**ክ**" while the same subject is represented as "+**ከ**" in the Ge'ez (**ቀተልከ** /you killed/). Likewise, the definite marker for Amharic and English are quite different in their representation. While it is a bound morpheme in Amharic, it is a word (free morpheme) in English. Most of the local languages under consideration falls into this group.

4 Parallel Corpus 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 languages paired with English. The collected data fall under the religious, historical and legal domains.

The religious domain include Holy Bible and different documents written in spiritual theme and collected from Jehovah's Witnesses (JW[1]), Ethiopicbible[2], Ebible[3] and Ge'ez experience[4] which are freely available websites.

The historical domain is from one source which is the handbook of Africa ("African Almanac"). The source is griped from admase ethiopia github[5].

The legal domain includes documents collected from Ethiopian constitution, Proclamation and Regulation documents which are available for different period of time and languages (Amharic,

[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`
[5]Corpus available at https://github.com/admasethiopia/parallel-text/

Tigrigna and Afan-Oromo aligned with English). The documents are taken from Ethiopian legal brief website.

Legal and historical domain data collected from sources specified above are available in text and pdf format. For the sources in pdf, a pdf miner tool is used for extracting texts.The contents in the pdf files are stored in multiple columns with a language per column. By using a Unicode range of characters, the contents in each column were extracted without distorting the sentence sequence. For the corpus in the religious domain, a simple web crawler was used to extract parallel text from targeted websites.

Python libraries such as requests and BeautifulSoup were used to analyze the structure of the website, extract texts and combine to 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 verse numbers of Bible in each language.

For the daily text which is published at Jehovah Witnesses (JW), we tried to use the date information to generate URL for each language. The page was requested to extract the data we are interested in. Finally, we organized and merged the data to a single UTF-8 text files for each language.

We could have all these domains only for a language pair Amharic-English. The Tigrigna-English and Afan Oromo-English corpora are in legal and religious (both bible and other religious collections) domains. The Wolaytta-English and Ge'ez-English language pairs are from the religious domain only. However, the Ge'ez-English corpius is only from Bible while the Wolaytta-English consists of Bible and other religious collections.

4.1 Preprocessing

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 manually. To produce parallel corpus there is a need to analyze the structure of collected raw data by applying different techniques.

During preprocessing the following tasks have been performed: character normalization, sentence tokenization and alignment.

4.1.1 Character Normalization

There are characters in Amharic that have similar roles and are redundant. Characters ሀ, ሐ and ኀ; ሠ and ሰ; አ and ዐ as well as ጸ and ፀ are variants along with their orders. These characters are used interchangeably. To avoid words with the same meaning from being taken as distinct words due to these character variants, we have replaced a set of characters with similar function into a single most frequently used character.

As a result of normalizing character variants in Amharic text , reduction in the number of word types (vocabulary) has been obtained. Table 1 presents the vocabulary reduction of training, development and testing dataset. As can be seen from the table, the vocabulary size reduced by 15.78 % for training, 10.53% for development and 11.95% for testing from a total of 40,726 sentence (132,723 Token of 628,474 type).

	Word Normalization		**Percentage reduction**
	Before	**After**	
Training	98,784	83,196	15.78 %
development	23,701	21,207	10.53%
Testing	24,142	21,258	11.95%

Table 1: Amharic text corpus before and after character normalization.

4.1.2 Sentence Tokenization and Alignment

Lines that contain multiple sentences in both source and target languages are tokenized. The team have 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 and target languages. The second criterion is mapping the sentence end marker in source and destination languages. For example, after sentence tokenization is applied the number of new sentences and sentence end marker of the source is compared with that of the target. If both criteria are fulfilled, the new sentence list is used as a parallel corpus.

For the English–Ge'ez[6] parallel corpus, the source language contain multiple verses in a single line while on the target side, each line contains a single verse. To align the two corresponding language pairs, we tried to merge verses to produce the desired parallel verse. In addition, removing unnecessary links, numbers, symbols and foreign texts in each language has been done.

4.2 Corpus Size and Distribution of tokens and vocabulary

The corpus has been analyzed to see the relationship between English and each one of the considered Ethiopian languages. As it can be seen from Table 1 to Table 5 and the corresponding Figures (Figure 1 to 5), there is a significant difference between the morphology of English and the Ethiopian languages. Due to this difference, the same number of sentences in these language pairs is tokenized into significantly different number of tokens and vocabularies in all the available domains.

The Figures clearly show that English vocabulary is much lower than vocabulary of all the considered Ethiopian languages. On the contrary, the English token is significantly higher than tokens of the Ethiopian languages. It is clear, therefore, that such differences between the languages in a language pair makes SMT difficult because it aggravates data scarcity and results into a weakly trained translation model. The morphological complexity of the Ethiopian languages also challenges the SMT towards them because it results into a poorly trained language model.

	History	Legal	Religion	
			Bible	Blog
English *	35,325	85,526	767,989	80,505
Amharic *	29,804	63,940	472,294	62,436
English **	8,420	9,029	39,113	9,838
Amharic **	10,560	12,779	93,001	16,383

Table 2: Distribution of Tigrigna and English text.

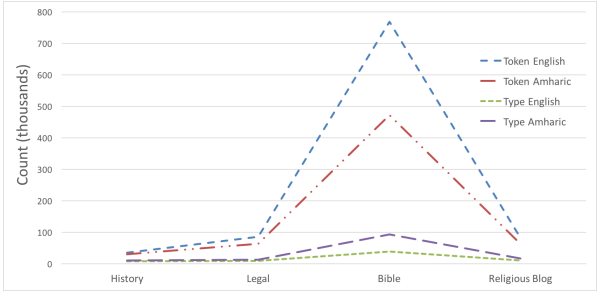

Figure 1: Comparison of Amharic-English SMT data

		Legal	Religion	
			Bible	Blog
Token	English	11,597	495,780	53,999
	Tigrigna	15,481	767,989	66,408
Type	English	2,989	81,674	13,494
	Tigrigna	2,286	39,113	8,818

Table 3: Distribution of Tigrigna and English text.

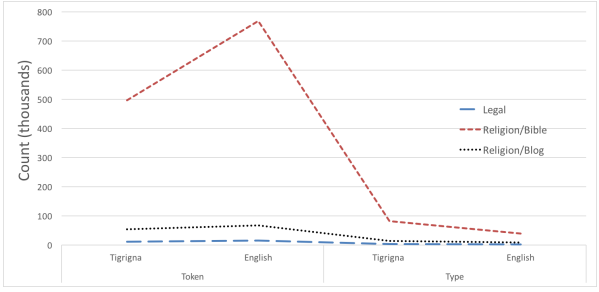

Figure 2: Comparison of Tigrigna-English SMT data

[6]English-Ge'ez parallel corpus available at https://www.ethiopicbible.com/amharic-bible-books

		Legal	Religion	
			Bible	Blog
Token	English	42,390	157,346	68,299
	Afan-Oromo	49,701	187,926	72,588
Type	English	7,819	19,844	10,110
	Afan-Oromo	7,819	13,659	9,320

Table 4: Distribution of Afan-Oromo and English text.

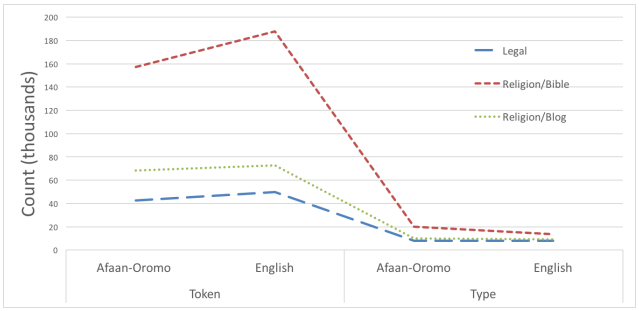

Figure 3: Comparison of Afan Oromo-English SMT data

		Religion	
		Bible	Blog
Token	English	468,122	41,041
	Wolaytta	700,321	59,754
Type	English	59,320	10,012
	Wolaytta	26,610	8,402

Table 5: Distribution of Wolaytta and English text.

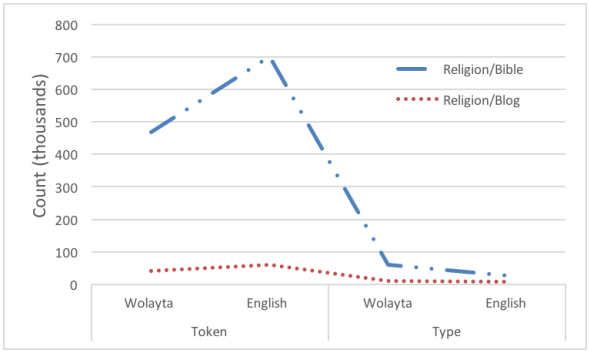

Figure 4: Comparison of Wolaytta-English SMT data

		Bible
Token	English	160,662
	Ge'ez	303,546
Type	English	33,894
	Ge'ez	15,260

Table 6: Distribution of Ge'ez and English text.

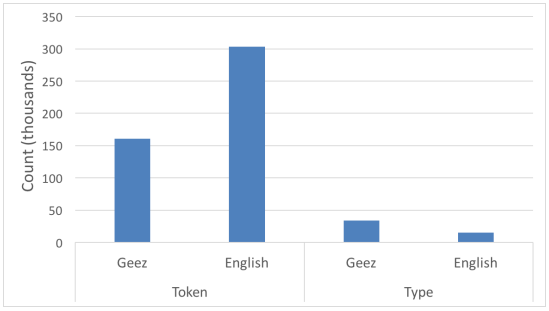

Figure 5: Comparison of Ge'ez-English SMT data

5 SMT Experiments and results

In this study, bi-directional SMT systems are developed to check the validity of the collected parallel corpora for English and the four Ethiopian languages.

5.1 Experimental setups

In the experimental setup, Moses is used along with GIZA++ alignment tool (Och and Ney, 2003) for aligning words and phrases. SRILM toolkit was used to develop language models using semi-automatically prepared corpora from the training and tuning corpora of target languages.

Table 7 shows the sentence length, the number of words, the total number of tokens and the average sentence length found in the corpora for the four Ethiopian languages with respect to English.

To carry out the experiments, each parallel corpus is divided into three partitions; 80% as a training set where a large subset of the whole corpus is used to train the language and translation models, 10% for tuning (useful to adjust the weights of the model combination) and 10% as a tes tset for evaluating the final bi-direction statistical machine translation system of each language pair.

Automatic metrics and subjective evaluation are the two most widely used techniques or methods for MT system evaluation. In this research, BiLingual Evaluation Under Study (BLEU)

		Sentence	Token	Type	Average word
Language Pairs	English	40,726	66,400	969,345	23
	Amharic		132,723	628,474	15
	English	35,378	50,217	849,878	19
	Tigrigna		98,157	561,376	14
	English	14,706	29,076	264,790	20
	Afan-Oromo		37,773	268,035	17
	English	30,232	35,012	760,075	21
	Wolaytta		69,332	509,163	14
	English	11,663	15,260	303,546	26
	Ge'ez		33,894	160,662	13

Table 7: Sentence and word distribution of Ethiopian languages and English text.

is used for automatic scoring. Table 8 presents the experimantal results of bi-directional English-Ethiopian languages SMT.

Language pair	BLEU	Language pair	BLEU
English-Amharic	13.31	Amharic-English	22.68
English-Tigrigna	17.89	Tigrigna-English	27.53
English-Afan Oromo	14.68	Afan Oromo-English	18.88
English-Wolaytta	10.49	Wolaytta-English	17.39
English-Ge'ez	6.76	Ge'ez-English	18.01

Table 8: Experimental results of bi-directional English-Ethiopian languages SMT

As shown in Table 8, the English-Amharic translation shows a BLEU score of 13.31 while the Amharic-English has a 22.68. Similarly, the English-Tigrigna and Tigrigna-English have BLEU scores of 17.89 and 27.53, respectively. Likewise, English-Afaan Oromo has a 14.68 BLEU and Afan Oromo-English has 18.88 BLEU score. In a similar way, the English-Wolaytta translation has BLEU of 10.49 while Wolaytta-English has 17.39. Finally, The English-Ge'ez and Ge'ez-English translation has BLEU score of 6.67 and 18.01, respectively.

Figure 6 presents summary of BLEU score registered for bi-directional English-Ethiopian languages using statistical approach.

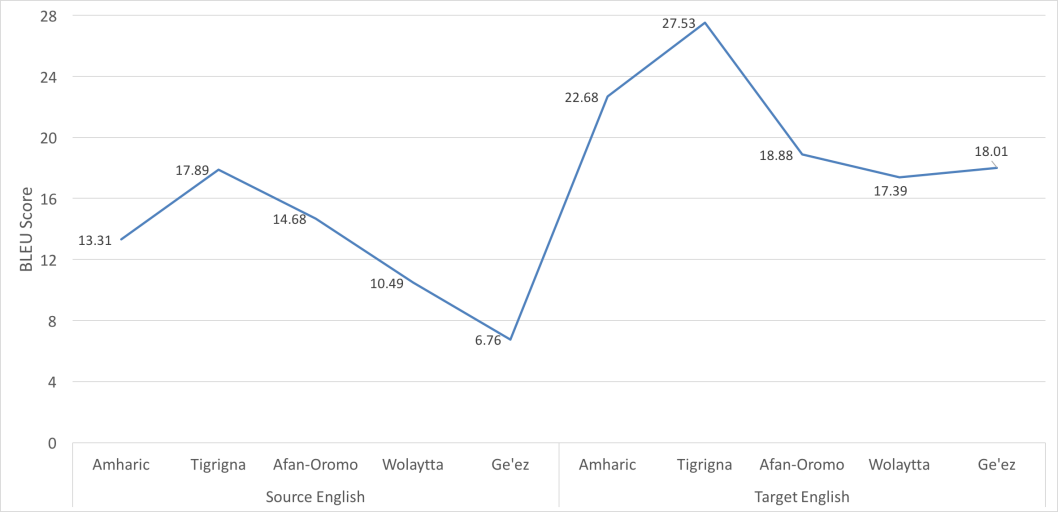

Figure 6: Comparison of Bi-directional English to Ethiopian language translation results

The BLEU score of Amharic-English translation system is lower than the Tigrigna-English translation system although the size of the Amharic-English parallel corpus is bigger than the Tigrigna-English one. This might be due to the number of domains considered in the corpora. The Amharic-English corpus covers all the three domains whereas the Tigrigna-English corpus is from only two domains.

Despite the size of the data, the English-Ethiopian languages SMT systems have less BLEU scores than that of Ethiopian languages-English ones. This is because of the fact that when the Ethiopian languages are used as a target language, the translation from English as a source language is challenged by many-to-one alignment. On the other hand, better performance is registered when the target language is English since the alignment is one-to-many taking each Ethiopian language as a source. In addition to this, the language model data favours the English language than that of Ethiopian languages due to the complexity of the morphology of these languages.

6 Conclusion and future work

This paper presents the attempt made in preparing standard parallel corpora for English and Ethiopian languages. The text data have been collected from the web in history, legal and religious domains. Then, the data are further pre-processed and normalized in preparing a bilingual parallel corpora for SMT task. Using the corpora, bi-directional statistical machine translation experiments have been conducted. The experimental results show that a translation from Ethiopian languages to English resulted in better BLEU score than that of the English to Ethiopian languages. The morphological richness of the Ethiopian languages greatly affect the performance of SMT specially when they are target languages.

To further see the impact, there is a need to conduct additional experiments with the objective of identifying an optimal one-to-many and many-to-one alignment when either of them used as a target language. Moreover, further research is needed to identify the exact reason behind the low performance of English to Ethiopian languages translation systems. Investigating the effect of domains on SMT performance is one of the future work we will work on.

References

Saba Amsalu and Sisay Fissaha Adafre. 2006. *Machine Translation for Amharic: Where we are.*, In proceedings of LREC 2006, pp. 47-50.

Philipp Koehn. 2009. *Statistical machine translation.*, volume 1. Cambridge University Press.

W.John Hutchins 1995. *Concise history of the language sciences: from the Sumerians to the cognitivists.*, volume 1. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, pp. 431-445

Tariku Tsegaye 2004. *English-Tigrigna Factored Statistical Machine Translation.*, MSc. Thesis, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.

Sisay Adugna Chala 2009. *English-Afaan Oromo Machine Translation: An Experiment Using Statistical Approach.,*. MSc. Thesis, School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia.

Eleni Teshome 2013. *Bidirectional English-Amharic Machine Translation: An Experiment Using Constrained Corpus.,*. MSc. Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

Jabesa Daba 2013. *Bi-directional English-Afaan Oromo Machine Translation Using Hybrid Approach,*. MSc. Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

Akubazgi Gebremariam 2013. *Amharic-Tigrigna Machine Translation Using Hybrid Approach,*. MSc. Thesis, Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia.

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

Sisay Fissaha Adafre. 2004. Adding Amharic to a Unification based Machine Translation System: An Experiment, ISBN: 9780820473314, Peter Lang GmbH.

Sisay Fissaha Adafre. 2004. *Adding Amharic to a Unification based Machine Translation System: An Experiment*, ISBN: 9780820473314, Peter Lang GmbH.

Michael Melese Woldeyohannis and Million Meshesha. 2017. *Experimenting Statistical Machine Translation for Ethiopic Semitic Languages : The case of Amharic-Tigrigna.*, International Conference on ICT for Development for Africa (ICT4DA) September 25–27, 2017 Bahir Dar, Ethiopia.

Gary F. Simons and Charles D. Fennig. . 2017. *Ethnologue: Languages of the World.* 20th Edition, SIL, Dallas, Texas.

John Hutchins. 2005. *The history of machine translation in a nutshell.*. Retrieved March, 2018, pages 1–5, 2005. URL http://www.hutchinsweb.me.uk/Nutshell-2005.pdf

Leslau, W. 2000. Alternation. *Introductory Grammar of Amharic.* Otto Harrassowitz, Wiesbaden.

Teferra, A. and Hudson, G. 2007. *Essentials of Amharic.* Rudiger Koppe Verlag.

Wakasa, M. 2008. *A Descriptive Study of the Modern Wolaytta Language.* University of Tokyo.

Mason, J. S. 1996. *Tigrigna grammar.* Tipografia U. Detti.

Yohannes, T. 2002. *A Modern Grammar of Tigrigna.* Tipografia U. Detti.

Griefenow-Mewis, C. 01. *A grammatical sketch of written Oromo.*, volume 16. Rüdiger Köppe.

Gasser, M. 2010. *A Dependency Grammar for Amharic.*, In Workshop on Language Resources and Human Language Technologies for Semitic Languages.

Gasser, M. 2011. *HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrigna.*, In Conference on Human Language Technology for Development. Alexandria, Egypt.

Dillmann, A. 1907. *Ethiopiac Grammer*, 24(11):503–512. Improved and enlarged by Karl Bezold, Translated by J.A. Crichton. London: William and Norgate.

Och, F.J. and Ney, H. 2003. *A systematic comparison of various statistical alignment models.*, 29.1 (2003): 19-51. Computational linguistics.