English-Ethiopian Languages Statistical Machine Translation

Solomon Teferra[1], **Michael Melese**[1], **Martha Yifiru**[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 SMT experiments. The BLEU scores of the bi-directional 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.

1 Introduction

The advancement of technology and the rise of the internet as a means of communication led to an ever increasing demand for NLP applications. One NLP applications which facilitates human-human communication is Machine Translation (MT). In the presence of high volume digital text, the ideal aim of MT 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 17^{th} century in corpus-based approach (Hutchins, 1995; Koehn, 2009). A corpus based approaches require parallel and monolingual corpora without deep linguistic analysis.

Furthermore, research in the area of MT for Ethiopian languages, which are under-resourced as well as 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 MT system (Sisay, 2004) and the other that investigated English-Amharic SMT (Mulu, 2017). Beside these, Michael and Million (2017) attempted a bi-directional Amharic-Tigrigna SMT experiment using different translation units.

African languages, which contribute around 30% (2139) of the world languages, highly suffer from lack of sufficient NLP resources which is true for Ethiopian language too (Simons and Fennig, 2017). However, a lot of written documents in the web are being produced in technological favored languages such as English. Due to unavailability of linguistic resources and since the most widely used MT approach is statistical, most of the researches have been conducted using SMT, which requires parallel and monolingual corpora. However, as there were no such corpora for SMT experiments, we have collected and prepared parallel corpora for English and Ethiopian languages considering Amharic, Tigrigna and Ge'ez from the Semitic, Afan-Oromo from the Cushitic and Wolaytta from Omotic families. This paper, therefore, describes an attempt made to collect and prepare English-Ethiopian languages corpora for SMT experiments.

2 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, 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 corpus is only from Bible while the Wolaytta-English consists of Bible and other religious collections.

After collecting the 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 sentence alignment.

3 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. To carry out the experiments, each parallel corpus is divided into three partitions; 80% as a training set, 10% for tuning and 10% as a testset for evaluating the final bi-directional SMT 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) is used for automatic scoring. Table 1 shows distribution of four Ethiopian languages with respect to English while Table 2 presents bi-directional English to Ethiopian language SMT evaluation result using BLEU score.

[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/

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

Table 1: Distribution of parallel corpus.

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

Table 2: Experimental results of bi-directional English-Ethiopian languages SMT

As shown in Table 2, 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 while Afan Oromo-English has 18.88. 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. 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.

4 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 SMT 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.