English-Afaan Oromo Statistical Machine Translation

Million Meshesha *million.meshesha@aau.edu.et*
School of information science
Addis Ababa University
Addis Ababa, Ethiopia

Yitayew Solomon *yitayewsolomon3@gmail.com*
Information technology
Metu University
Metu, Ethiopia

Abstract

Statistical machine translation (SMT) is an approach that mainly uses parallel corpus for translation and its performance is dependent on effectiveness of alignment of source and target languages. This study explores the effect of word, phrase and sentence levels of alignment on English-Afaan Oromo statistical machine translation. We used GIZA++, Anymalignment and hunalign for word level, phrase level and sentence level alignment, respectively. Experimental result shows that **27%** BLUE score is recorded at phrase level alignment with maximum phrase length of 16. The Syntactic structure sensitivity of the alignment tool and the challenge of word correspondence variation in the two languages needs further investigation.

Keywords: Statistical Machine Translation, Afaan Oromo Language, Word Correspondence Alignment.

1. INTRODUCTION

Natural language is one of the fundamental aspects of human behavior and a crucial component in our lives. It is a tool for communicating all around the world. Natural language processing (NLP) can be described as the ability of computers to generate and interpret natural language [1]. Machine translation is the application of computers to the task of translating text and speech from one natural (human) language such as English to another human language such as Afaan Oromo [2]. Afaan Oromo is one of the languages of the Low land East Cushitic within the Cushitic family of the Afro-Asiatic Phylum [3, 4]. It is also one of the major Languages spoken in Ethiopia. According to Gene [5] and Hamid [6], Afaan Oromo is the third most widely spoken language in Africa after Arabic and Hausa. Oromo language, also referred to as Afaan Oromo or Oromiffaa has more than 20 million speakers which is the second most widely spoken indigenous language in Africa [7]. More than two-thirds of the speakers of the Cushitic Languages are Oromo or speak Afaan Oromo, which is also the third largest Afro-Asiatic language in the world [7]. In spite of its usage, as a vernacular, the language is widely spoken in the Horn of Africa [7].

The typological facts about cross-linguistic similarities and differences that were studied include word order of noun, verb and objects in simple declarative clauses [8]. For example, in English, a simple declarative sentence is in Subject-Verb-Object (SVO) order while in Afaan Oromoo it is in Subject-Object-Verb (SOV) order. Yet another typological fact is the word order of noun and adjective in the two languages. For example, in English, nouns follow adjectives (as in excellent student) while in Afaan Oromoo the reverse is true (as in bartaa ciimaa). Here ciimaa is an adjective and it means ‘excellent’ and bartaa is a noun and it means ‘student’. The researcher believes that these cases have something to do in the tasks of word alignment, language modeling, translation modeling and decoding.

MT has different approaches, including rule based, corpus based and hybrid [2]. Rule-Based Machine Translation, also known as Knowledge-Based MT, is a general term that describes machine translation systems based on linguistic information about source and target languages. Corpus-based MT approach, also referred as data driven machine translation, is an alternative approach for machine translation to overcome the problem of knowledge acquisition problem of rule based machine translation. Corpus Based Machine Translation uses, a bilingual parallel corpus to obtain knowledge for new incoming translation. By taking the advantage of both corpus based and rule-based translation methodologies the hybrid MT approach is developed, which has a better efficiency in the area of MT systems [3].

Machine translation has its own challenges and still an active research area [8]. The challenges are translation of low-resource language pairs, translation across domains, translation of informal text, translation of speech and translation form/to morphologically rich languages.

Machine translation (MT) systems have been developed by using different methodologies and approaches for pairs of foreign languages [9, 10]. Most study for local languages are more focused on Amharic [1, 11] and Afaan Oromo languages [12, 13]. Sisay [12], conducted an experiment on English-Afaan Oromo language pairs by using statistical MT approach. Another experiment which was done by Jabesa [13], explores a bidirectional English-Afaan Oromo machine translation that compares rule based with statistical machine translation (SMT) approach.

The main challenge both researchers emphasized was the alignment quality of the prepared dataset due to the unavailability of well-prepared corpus for the statistical machine translation task. This shows the need for undertaking further study to identify an optimal alignment for the prepared Afaan Oromo-English parallel corpus. It is therefore the aim of this study to identify optimal alignment for English-Afaan Oromo statistical machine translation by studying the structure of both target and source languages.

2. ALIGNMENT CHALLENGE OF ENGLISH – AFAAN OROMO LANGUAGES

Afaan Oromo and English have differences in their syntactic structure. In Afaan Oromo, the sentence structure is subject-object-verb (SOV), where the subject comes first, followed by the object and the verb comes at the end of the given sentence. For example, if we take Afaan Oromo sentence “caalaan midhaan nyaate”, “caalaan” is the subject, “midhaan” is the object and “nyaate” is the verb of the sentence. In case of English, the sentence structure is subject-verb-object. For example, if the above Afaan Oromo sentence is translated into English it will be “caalaa ate food” where “caalaa” is the subject, “ate” is the verb and “food” is the object [12]. This difference in the syntactic structure affects effectiveness of the alignment task during text translation from source language to target language.

Alignment plays a critical role in statistical machine translation by mapping source sentence to target sentence [3]. However, automatic alignment of parallel sentence pairs is not a simple task. For most parallel texts, choosing the sentences in one natural language to be the translation of another language is a challenging activities. Words may have different level of alignments, such as one to one, one to many, many to one and/or many to many. This makes alignment of words difficult. Figure 1 below shows sample alignment properties of English and Afaan Oromo text from both direction.

As shown in Figure 1, there are different levels of alignments observed in a given parallel texts taken from English and Afaan Oromo languages. This is because of differences in the length of sentence constructs of the two languages based on concept mapping from English to Afaan Oromo, vis-a-vis. This non-linear correspondence between the two languages has a great effect in the alignment process for designing a statistical machine translation.

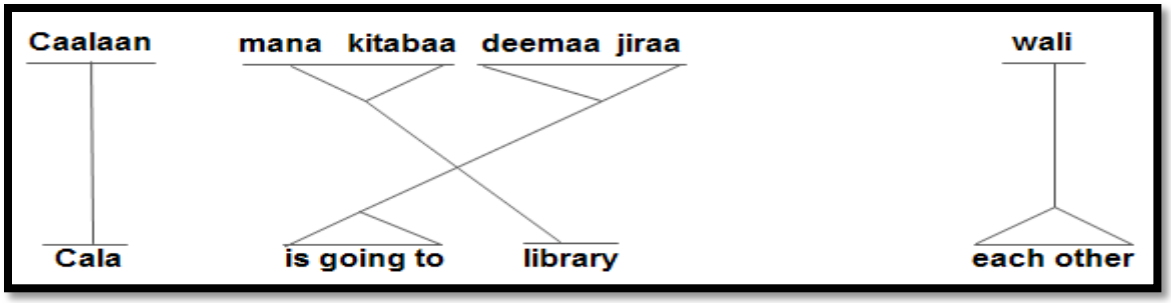

FIGURE 1: Alignments of English and Afaan Oromo Sentences.

3. METHODOLOGY

This study follows experimental research which requires data preparation, tool selection for constructing translation model and evaluation of the performance of the model.

3.1 Data Preparation

To perform the experiments, the data set or corpus was collected from Ethiopian criminal code, Ethiopian constitution, Oromia Regional State Duties and Responsibilities and Holy Bible. The reason to select these sources of data for corpus preparation is, because, the data is easily accessible from the web and they are parallel corpus which is suitable for the SMT task.

We performed data cleaning during preprocessing stage to make the data set ready for alignment and experimentation. The size of the corpus used for the experiments is 6400 sentences, prepared from the above mentioned online sources. We used 19300 and 12200 sentences as a monolingual corpora for creating English and Afaan Oromo language models, respectively.

3.2 Approaches

Statistical approach for machine translation is economically wise. Which doesn't require linguist professionals for corpus preparation, the translation process is done by using corpus. It is especially suitable for under resourced languages such as Afaan Oromo language. The basic tools we used for accomplishing the machine translation task is Moses for Mere Mortal; freely available open source software which is used for statistical machine translation. This software integrates different toolkits such as IRSTLM for language model, Decoder for translation. We used MGIZA++ for word alignment, Anymalign for phrase level alignment and hunalign for sentence level alignment in order to align the prepared corpus at different levels and explore their effect on the performance of SMT using BLUE score metrics.

4. THE PROPOSED SMT SYSTEM

Figure 2 depicts the architecture designed for experimenting English-Afaan Oromo statistical machine translation.

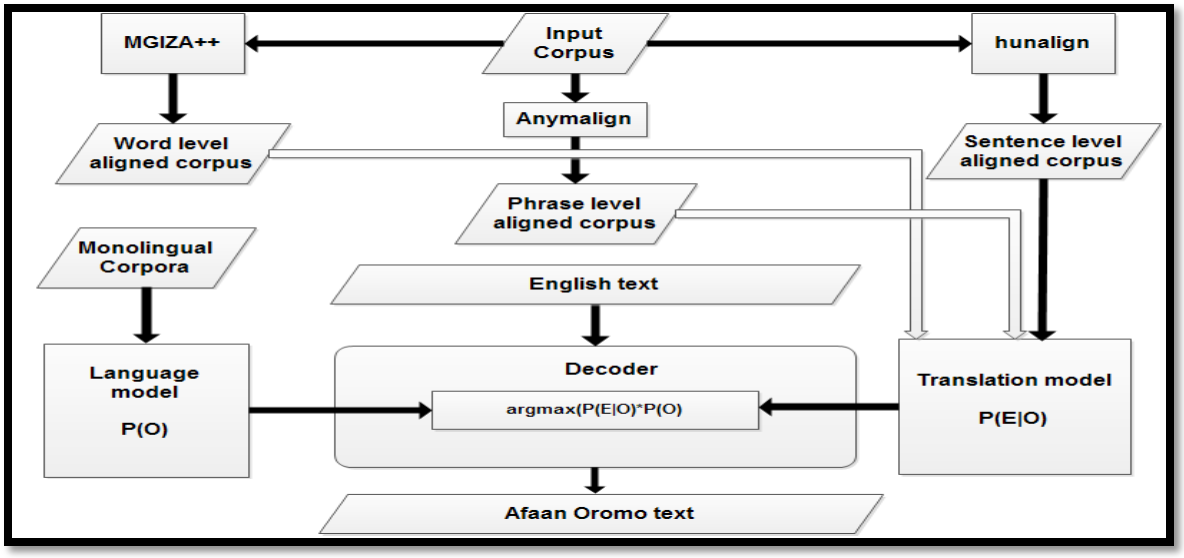

FIGURE 2: Architecture of The Proposed System.

The system accepts parallel corpus of English to Afaan Oromo and align at word, phrase and sentence levels using MGIZA++, Anymalign and hunalign respectively. The output of the alignment tool is used for creating translation model. For language model we used monolingual corpora of each language. While the language model computes prior probability distribution of English, P(E) and Afaan Oromo, P(O) languages, translation model calculates likelihood probability distribution, P(E/O)-the probability of occurrence of English text given Afaan Oromo text.

The decoder uses prior probabilities and likelihood probabilities to search for the shortest path in an implicit graph [1]. A decoder searches for the best sequence of transformations that translates source sentence in English to the corresponding target Afaan Oromo language. Mathematically, the decoder determine the maximum posterior probability for performing the translation from English to Afaan Oromo language.

$$P(O/E) = \text{argmax}_O\ P(E/O) * P(O)$$

5. EXPERIMENTATION AND PERFORMANCE ANALYSIS

In this study a three phase experiment is conducted using the aligned corpus at word level, phrase level and sentence level with phrase length from 1 to 4 words, 5 to 16 words and 17 to 30 words, respectively. The logic behind conducting such experiments are to measure the effect of different levels of phrase length aligned corpus on the performance of English to Afaan Oromo statistical machine translation. Accordingly experimental result is presented in the table 1 below.

Alignment	Phrase length	BLUE score	Time taken
MGIZA++	1 to 4	21%	14s
Anyalign	5 to 16	**27%**	12s
Hunalign	17 to 30	18%	17s

TABLE 1: Summary of Experimental Result.

The translation experiments conducted at different levels of alignment from English to Afaan Oromo text shows that the performance registered at maximum phrase length 16 is better than the other experiments with improvements of performance by 6% and 9% as compared to maximum phrase length of 4 and 30 respectively. The result confirms that phrase level alignment is better than word level and sentence level alignment. This is because most of the correspondence between English and Afaan Oromo language is word to phrase. This means that a combination of multiple words in Afaan Oromo have single word meaning in English; for example, " Mana kitabaa → Library".

In this study we found that, for designing English to Afaan Oromo SMT with a better performance the alignment level needs due attention, as word correspondence is not only one to one rather it includes one to many, many to one and many to many. Also the observed difference in the syntactic structure of the two languages, where English language follows Subject Verb Object (**SVO**) but, Afaan Oromo construct sentences with Subject Object Verb (**SOV**), increase the complexity of English to Afaan Oromo text translation. This creates an added complexity during the alignment process since the alignment tool is expected to go in non-linear fashion to identify word correspondence.

6. CONCLUSION AND RECOMMENDATION

In this study an attempt is made to apply English to Afaan Oromo statistical machine translation. As a matter fact, statistical machine translation and alignment of the corpus have a strong relation because, in order to translate text following SMT approach the system has to learn from properly aligned corpus to construct translation model. This paper cover word level, phrase level and sentence level alignment by considering the structure of the source language English and target language Afaan Oromo.

The design process of English-Afaan Oromo statistical machine translation involves collecting parallel corpus from the available online resources. The corpus is cleaned and aligned to create parallel corpus for training and creating translation model. Moses for mere mortal used for translation process which integrate all necessary tools for machine translation such as IRSTLM, MGIZA++ and decoder.

Experimental results shows that a BLUE score of 27% achieved at maximum phrase length 16 in translating English text to Afaan Oromo language. In this study a promising result is registered. Phrase length and alignment verities are major challenges for the current study.

Our investigation found that phrase length and alignment verities that happened because of syntactic structure and differences in word correspondence needs to be controlled to improve the performance of the translation system for the two languages.

7. ACKNOWLEDGMENT

Above all I would like to thank the almighty God, who gave me the opportunity and strength to achieve whatever I have achieved so far. I would like to express my gratitude to all the people who supported and accompanied me during the progress of this work.First, I would like to express my deep-felt gratitude to my advisor, Dr. Million Meshesha, whose excellent and enduring support shaped this work considerably and made the process of creating this work an invaluable learning experience.

I want to thank Dr. Marta Yifru for helped me by sharing her experience on title selection before the beginning of the work and Sisay Adugna helped me by sharing his experience on his previous work on machine translation. I also wants to thank tool developer used in this study Maria Jose Machado and Hilario Leal Fontes (Moses for Mere Mortal), Pavel Vondericka (Inter Text editor 'hunalign'), and Adrien Lardilleux and Yves Lepage (Anymalign).

Finally I want to thank my friends and colleagues (Zebider Birhane, Ramata Mossisa, Mesay Wana and Haile Michael Kafiyalew), who helped me by reading the work and gives constructive

comment and Bewunetu Dagne helped me by supporting on the installation of the tools used for this study.

8. REFERENCES

[1] E. Teshome, "Bidirectional English-Amharic machine translation An Experment based on constriented corpus,"Msc thesis Addis Ababa university, Adis ababa Ethiopian, 2013.

[2] A. Mouiad , O. Nazlia and S. M. Tengku , "Machine Translation from English to Arabic," International Conference on Biomedical Engineering and Technology, vol. 11, pp. 95-99, 2011.

[3] M. Bulcha, "Oromo Writing," Nordic Journal of African Studies, pp. 36-59, 1995.

[4] G. B. Gene , Students in Ancient oriental civilayzation No.60, S. leslie and U. G. Thomas, Eds., chicago: university of chicago, 1982.

[5] D. Fufa, "Indigenous Knowledge of Oromo on Conservation of Forests and its Implications to Curriculum Development: the Case of the Guji Oromo," Addis ababa, 2013.

[6] M. Hamid , Oromo dictionary: English-Oromo, Atlanta: Sagalee Oromoo, 1995.

[7] M. Hundie, "lexical standardization," Addis ababa, 2002.

[8] A. Lopez and M. Post, "Beyond bitext: Five open problems in machine translation," Human Language Technology Center of Excellence, Vol. 5, No 3. 2011.

[9] H. Somers, "Machine translation latest devlopments," in Readings in Machine Translation, . N. Sergei, S. Harold and W. Yorick , Eds., manchister, MIT Press, 2003, pp. 513-528.

[10] S. Holger , F. Jean-Baptiste and S. Jean , "First steps towards a general purpose French/English statistical machine translation system," Association for Computational Linguistics, pp. 119-122 , 19 june 2008.

[11] M. G. Teshome and B. Laurent , "Preliminary experiments on english-amharic statistical machine translation," pp. 36-41, 2012.

[12] S. Adugna, "English-Oromo Machine Translation: An Experiment Using a Statistical Approach," Msc thesis Addis Ababa University, Adis Ababa Ethiopia , 2009.

[13] J. Daba, "Bidireactional English – Afaan oromo Machine translation using hybrid approach," Msc thesis Addis Ababa University, Adis Ababa Ethiopia, 2013.

[14] J. Daniel and M. H. James , Speech and Language Processing: An introduction to natural language processing, 2008.