Addis Ababa University

Collage of Natural and Computational Science

School of Information Science

Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Information Science

By:

Yitayew Solomon

(syite@ymail.com)

Advisor: Million Meshesha (PhD)

Addis Ababa, Ethiopia

June, 2017

Dedication

I dedicate this work to my mother “Ayinalem Mersha”.

Look up to the sky

Now tell me what you see

A cloud, the moon, possibly the sun

Many answer there will be

When I look up to the sky

I will tell you what I see

I see my mother

And she’s looking back at me!!!

Addis Ababa University

Collage of Natural and Computational Science

School of Information Science

Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation

Signature for Approval

Name	Signature	Date
Million Meshesha (PhD) , Advisor	_____	_____
Marta Yifru (PhD), Examiner	_____	_____
Wondwossen Mulugeta (PhD), Examiner	_____	_____

Declaration

I declare that this research is my original work and has not been presented for a degree in any university, and that all sources of material used for the research have been properly acknowledged.

Declared by:

Name: Yitayew Solomon

Signature: _____

This research has been submitted for Examination with my approval as university advisor.

Name: Million Meshesha (PhD), Advisor

Signature: _____

Date: _____

Addis Ababa, Ethiopia

June, 2017

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.

Abstract

Statistical machine translation is an approach that mainly use parallel corpus for translation, in which parallel corpus alignment of the given corpus is crucial point to have better translation performance. Alignment quality is a common problem for statistical machine translation because, if sentences are miss aligned the performance of the translation processes becomes poor. This study aims to explore the effect of word level, phrase level and sentence level alignment on bi-Directional Afaan Oromo-English statistical machine translation.

In order to conduct the study the corpus was collected from different sources such as criminal code, FDRE constitution, Megleta Oromia and Holly Bible. In order to make the corpus suitable for the system different preprocessing tasks applied such as true casing, sentence splitting and sentence merging has been done. A total of 6400 simple and complex sentences are used in order to train and test the system. We use 9:1 ratio for training and testing respectively. For language model we used 19300 monolingual sentence for English and 12200 for Afaan Oromo. For the purpose of the system we used Mosses for Mere Mortal for translation process, MGIZA++, Anymalign and hunalign tools for alignment and IRSTLM for language model. After preparing the corpus different experiments were conducted.

Experiment results shows that better performance of **47%** and **27%** BLUE score was registered using phrase level alignment with max phrase length 16 from Afaan Oromo-English and from English-Afaan Oromo translation, respectively. This depicts an improvement of on the average 37 % accuracy registered in this study. The reason for this score is length of phrase level aligned corpus handle word correspondence. This depicts that alignment has a great effect on the accuracy and quality of statistical machine translation from Afaan Oromo-English and the reverse.

During machine translation alignment of a text of multiple language have different correspondence, one-one, one-many, many-one and many-many alignment. In this study, many-many alignment is a major challenge at phrase level that needs further investigation.

Key word: SMT; word level alignment; phrase level alignment; sentence level alignment; Afaan Oromo.

Table of Contents

List Of tables

List of figures

List of abbreviation

ALPAC – Automatic language processing Advisory committee

Anymalign – Any multi lingual aligner

BLUE – Bilingual Evaluation Understudy

DMT – Direct machine translation

EASMT – English – Amharic statistical machine translation

EBMT – Example based machine translation

FDRE – Federal democratic republic of Ethiopia

MMM – Mosses for mere mortal

MT – Machine translation

RBMT – Rule based machine translation

SL – source language

SMT – Statistical machine translation

TL – target language

CHAPTER ONE

Introduction

1.1 Background

Human language, whether written or spoken, is a fundamental part of human communication. 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 language [2]. Machine translation has different advantages; among them the following are common [1]: one of the advantage is Confidentiality. Since people use machine translation systems to translate their private information, people communicate only with the system (MT) than other individuals, as a result, the privacy of the individuals are protected. The second advantage is fast translation. By using machine translation system it is possible to save time while translating large texts even paragraph or document in short period of time. The third one is universality. Usually a human translator translate the meaning of the text in their own context. This may bias the meaning of the text; but, in case of machine translation a text will be translated with the same meaning anywhere and everywhere, this makes machine translation universal.

MT approaches includes 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. Statistical analysis techniques are applied to create models whose parameters are derived from the analysis of bilingual text corpora. Example-based machine translation (EBMT) is characterized by its use of bilingual dictionary with parallel texts as its main knowledge, in which translation by correlation is the main idea. 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 [4]. One of the challenge is translation of low-resource language pairs. This is the scarcity of data covers most of the world's language pairs. The other is translation across domains. Translation systems are not strong across different types of data, performing poorly on text whose underlying properties differ from those of the system's training data. The third challenge is Translation of informal text. People want to read blogs, social media, forums, review sites, and other informal content in other languages for the same reasons they read them in their own. However, informal data translation are scarce. Further challenge is translation into morphologically rich languages. Most MT systems will not generate word forms that they have not observed, a problem that pervades languages like Amharic and Afaan Oromo. Further challenge is Translation of speech. Much of human communication is oral. Even ignoring speech recognition errors, the substance and quality of oral communication differs greatly from that found in most cases.

According to [5], an important new development for MT in the last decade has been the rapid progress that has been made towards developing speech to speech machine translation. Once thought simply too difficult, improved speech-analysis technology has been coupled with innovative design to produce a number of working systems, albeit still experimental, which suggest that this may be the new growth area for MT research. There are two process of translations that are uni-directional and bi-directional process. Uni-directional works only in one direction, which is first the system (language model and translation model) train by using the data set in one direction from source to target language, and the translation process also done in one direction from source to target language but not the revers. In bi-directional, the system (language and translation model) is trained in both direction and the translation process also done in both direction from source language to target language and form target language to source language.

1.2 Statement of the problem

English is a language that is widely spoken on different parts of the world. Most of the materials, software or other published literatures are written in English. Afaan Oromo language is one of language spoken in Ethiopia, it is obvious that both Afaan Oromo and English speakers need the data or documents written in English or Afaan Oromo and they also need to communicate with each other.

According to the Web Characterization Project of the Online Computer Library Center (www.oclc.org), there are plentiful documents in English on the Internet. This collections are accessed by different people around the world. For purpose of research, in order to develop their knowledge and to share information. However, lack of English language knowledge creates a problem of utilizing these collection. We believe that studying how to make these documents available in local languages (such as Afaan Oromo) is vital in order to access valuable information from the collection. Therefore, machine translation plays an important role to handle language barriers between peoples and documents who want to access them.

Machine translation (MT) systems have been developed by using different methodologies and approaches for pairs of foreign languages [5, 6]. Most study for local languages are more focused on Amharic [1, 7] and Afaan Oromo languages [8, 9]. Sisay Adugna [8], conducted an experiment on English-Afaan Oromo language pair by using statistical MT approach. Another experiment which was done by Jabesa Daba [9], a "bidirectional English-Afaan Oromo machine translation using hybrid approach" that combines both rule based approach and statistical machine translation (SMT) approach. The BLUE score of both experiments ranges from 17% to 37% [8, 9]. The main reason cited by the researchers for the poor performance was the alignment quality of the prepared data due to the unavailability of well-prepared corpus for the machine translation task. This shows the need for undertaking further study to identify an optimal alignment for the prepared corpus used for training and testing.

Therefore, the aim of this study is to experiment on proper alignment quality of the corpus based on the structure of the source and target language using large corpus so as to enhance the performance of SMT.

To this end this study attempts to address the following research questions:

- What is the optimal alignment to use for statistical machine translation?
- To what extent the selected alignment improves the performance of statistical machine translation?

1.3 Objective of the study

1.3.1 General objective

The general objective of this research is to explore an optimal alignment for bi-directional English-Afaan Oromo SMT.

1.3.2 Specific Objectives

Specific objectives of this research are as follows:

- ✓ To review different approaches used in machine translation.
- ✓ To identify the syntactic relationship between English and Afaan Oromo languages.
- ✓ To explore different tools used to align corpus.
- ✓ To collect English-Afaan Oromo parallel corpus for training and testing purpose.
- ✓ To prepare suitable aligned corpus for word level, phrase level and sentence level experiments
- ✓ To construct a prototype for bi-directional English-Afaan Oromo, statistical machine translation.
- ✓ To evaluate the performance of the prototype.

1.4 Scope and limitation of the Study

Bi-directional English-Afaan Oromo, statistical machine translation is designed to translate a sentence written in English text into Afaan Oromo text and vice versa. In this research, speech to speech translation, text to speech translation and speech to text translation are not included in the study.

As we try to indicate in the statement of the problem the main focus of this research is to explore an optimal alignment for better performance of statistical machine translation from Afaan Oromo-English and vice versa.

The source of the data set include FDRE criminal code, FDRE constitution, Megeleta oromia and Holy Bible of English and Afaan Oromo version and simple sentences, because, these sources are easily available and they are parallel corpus which is suitable for SMT. To conduct the research we follow statistical MT approach, which involves preparing parallel corpus for both target and source language, aligning the prepared parallel corpus, using aligned parallel corpus to train the system in both direction and the finally performing a bi-directional machine translation from source to target language and from target to source language.

Because of unavailability of standardized corpus (corpus ready for MT research purpose) and balanced corpus(in terms of discipline) the data set prepared in this study focus on sources that are parallel textual data, as a result of which most of the data we used for training and testing are from legal document.

1.5 Significance of the Study

The rate of machine translation is exponentially faster than that of human translation [10]. The average human translator can translate around 2,000 words a day. One should note that the output of machine translation is not in its final useable form right away, but in certain scenarios it can be quite useful. Even when adding a post-editing step, machine translation takes a fraction of time that human translation takes. In relation with this the main significance of this research work are the following; the first one is it helps for individuals and organizations who works on translation manually to facilitate the translation process by using this system. The second importance is it solves language barriers between individuals in order to read and understand different publications. The third importance is it helps for designing cross-language information retrieval to translate the query pose by the users. The fourth importance is reaching under resourced language; by translating publications example from English to Afaan Oromo it is possible to address information need of Afaan Oromo language speakers.

1.6 Methodology of the study

Research methodology is a way to systematically solve the research problem [11]. It may be understood as a science of studying how research is done scientifically. The advantage of knowing the methodology of the study before doing the Experiment is in order to reason out what, how and

why the methods or the techniques are selected for the Experiment in order to know the risks for conducting the research in detail.

1.6.1 Research design

In order to conduct the research we follow experimental research design because, to explore an optimal level of alignment for better performance of statistical machine translation, different experiments are conducted. Experimental research [12] investigates the possible cause-and-effect relationship by manipulating independent variables to influence the dependent variable(s) in the experimental group, and by controlling the other relevant variables, and measuring the effects of the manipulation by some statistical means. Steps in Experimental Research include the following [12]; the first step is, devise alternative hypotheses. The second step is crucial experiments with alternative possible outcomes, each of which exclude one or more possible hypotheses, Experiment. The third step is Conduct the experiment, get a clean result.

1.6.2 Data collection

To perform the experiments, the data set or corpus was collected from FDRE criminal code, FDRE constitution; Megeleta Oromia, Holy Bible see the URL of sources on appendix [1] and simple sentences adapted from [8, 9]. The reason to select these sources of data for corpus preparation is, because, it is easily accessible from the web and they are parallel corpus which is suitable for SMT easily.

Size of the corpus for the experiment is 6400, prepared from the above mentioned source of corpus. A great effort is deployed to enhance the size of the corpus that was used in the previous studies conducted on this area [8, 9] which uses from 3000-4000 sentence. In terms of discipline, the data set taken 2000 from FDRE constitution, 2400 from FDRE criminal code, 700 from Megeleta Oromia, 600 from Holly Bible and 700 simple sentences adapted from [8,9]. The reason why we select more corpus from FDRE constitution is because of the availability of large amount of textual data with more coverage of the domain. We used 19300 and 12200 monolingual corpora for language model for English and Afaan Oromo languages respectively which is prepared from above mentioned source of corpus.

In order to sample corpus from these sources our basic criteria is the coverage of the contents and the accessibility of sources. Based on this criteria we sample 400 articles from 865 articles of

criminal code, 50 articles from 106 articles of FDRE constitution, whole document (26 pages) of Megeleta Oromia and from bible 28 chapter of St Matthew.

1.6.3 Approach and tools used for the study

Machine translation has different approaches such as, example based approach, and rule based approach, statistical approach and hybrid approach. Statistical approach is economically wise i.e. doesn't need linguist professionals, the translation process is done by only from parallel corpus and also recommended by different researchers [3] because, it is current research area for machine translation for this reason we used statistical approach for this study.

The basic tools used for accomplishing the machine translation task is Moses for Mere Mortal; free available open source software which is used for statistical machine translation and integrates different toolkits which used for translation purpose such as IRSTLM for language model, Decoder for translation, MGIZA++ for word alignment.

Since the aim of the study is identifying an optimal alignment for enhancement of the performance of SMT Hunalign; used for sentence level alignment in order to align the prepared corpus at sentence level. Anymalign (Any multi lingual aligner); used for phrase level alignment of prepared corpus which is written by python, and MGIZA++ used for word level alignment. These three alignment tools used in our study because, they are alignment tools which used in SMT research for alignment purpose and it goes with our objectives of the study.

1.6.4 Evaluation procedure

Machine translation systems are evaluated by using human evaluation method or automatic evaluation method. Since human evaluation method is time consuming and not efficient with respect to automatic evaluation method, we used BLEU score metrics to evaluate the performance of the system, which is automatic evaluation method.

Bilingual Evaluation Understudy (BLUE) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's translation output and that of a human translated output.

If the machine translation output closer to human translation output it is considered as better translation, this is the basic idea behind BLEU [13]. BLEU was one of the metrics to achieve a

high correlation with reference translation, and remains one of the most popular automated and inexpensive metrics used in different researches for evaluation purpose.

In order to evaluate the performance of the prototype first we prepare the translated text by the system and second human translated text which is used as reference translation, by using these two texts BLUE score metric evaluate the performance of the system.

1.7 Thesis organization

This thesis is organized in to six chapters, the first chapter discuss about introduction, statement of the problem, objective of the study, scope and limitation of the study, methodology followed including research design, data collection, approach for the study and MT Evaluation procedure.

The second chapter deals with literature review which focus on approach of machine translation, alignment and the effects of alignment on statistical machine translation, and different tools used for corpus alignment and related works related with this study.

The third chapter deals with over view of Afaan Oromo language and its relationship with English language and discussion of alignment challenge between English Language and Afaan Oromo language.

Chapter four discuss about designing processes of the prototype including, corpus preparation, types of corpus used for the study, corpus alignment, and briefly discuss about the proto type of the system. Chapter five deals with Experiment of the study which include different experiments and the results of the experiments with interpretation of findings. The last chapter is chapter six deals about conclusion of the findings and recommendations for further works.

CHAPTER TWO

Literature Review

2.1 Overview of machine translation

The history of machine translation is traced from the pioneers and early systems of the 1950s and 1960s, the impact of the ALPAC report in the mid-1960s, the revival in the 1970s, the appearance of commercial and operational systems in the 1980s, research during the 1980s, new developments in research in the 1990s, and the growing use of systems in the past decade resulted to the birth of machine translation [14].

2.2 Machine translation

The term machine translation refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-line documents, remote terminology databanks, transmission and reception of texts, etc. [15]. Machine Translation, as it is generally known is the attempt to automate all, or part of the process of translating from one human language to another [16].

2.3 Why machine translation?

In the modern world, there is an increased need for language translations owing to the fact that language is an effective medium of communication [3]. The demand for translation has become more in recent years due to increase in the exchange of information between various regions using different regional languages. Accessibility to web document in other languages, has been a concern for information Professionals and other individuals or organizations who want to satisfy their information need.

2.4 Process of machine translation

A machine translation (MT) system, first analyses the source language input and creates an internal representation [3]. This representation is manipulated and transferred to a form which is suitable for the target language. Then at last output is generated in the target language. On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed.

2.5 Machine Translation Approaches

Machine translation approach can be classified according to the methodology. There are two main approaches: the rule-based approach and the corpus-based approach [3]. In the rule-based approach, human experts sets rules to describe the translation process, so that a huge amount of input from human experts (linguist professionals) is required. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combination of the two approaches gave birth to the Hybrid Machine Translation Approach.

2.5.1 Rule-Based Machine Translation Approach

Rule-Based Machine Translation (RBMT), also known as Knowledge-Based Machine Translation, is a general term that describe machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively [3]. Having input sentences, an RBMT system generates them to output sentences on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a real translation task.

RBMT methodology applies a set of linguistic rules in three different phases [3]: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation as shown in figure 2.1 below:

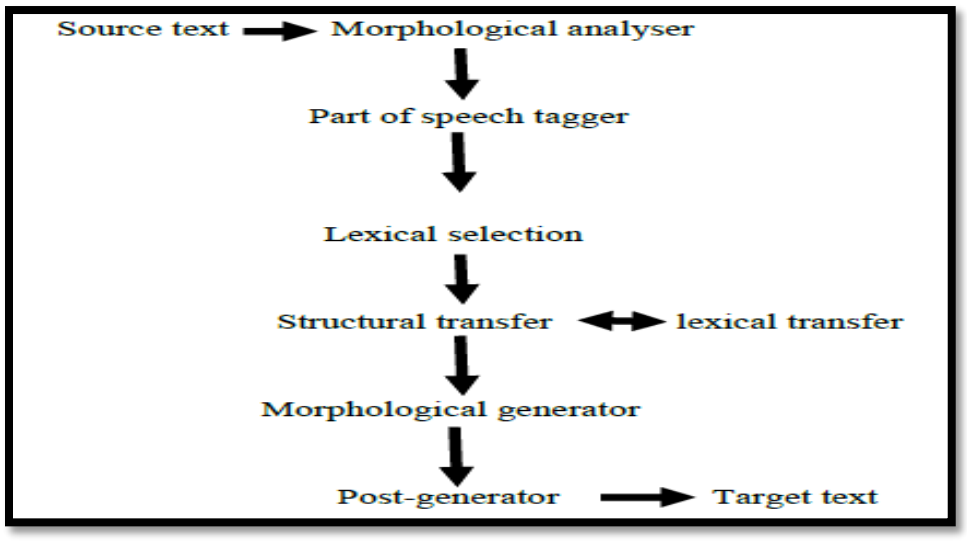

Figure 2.1: Architecture of rule based machine translation

The following are the shortcomings that are associated with RBMT approach [3]; Insufficient amount of good dictionaries, building new dictionaries is expensive, some linguistic information still needs to be set manually, hard to deal with rule interactions in big systems and ambiguity, and Failure to adapt to new domains.

2.5.1.1 Approaches of RBMT

There are three different approaches under the rule-based machine translation Approach [3]. They are Direct, Transfer-Based and Interlingua Machine Translation Approaches. They differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning between the source and target languages.

Direct Machine Translation (DMT) Approach: DMT approach is the oldest and less popular approach. Direct translation is made at the word level. Machine translation systems that use this approach are capable of translating source language directly to target language. Direct translation systems are basically bilingual and uni-directional. This approach needs only a little syntactic and semantic analysis. DMT is a word-by-word translation approach with some simple grammatical adjustments.

Inter-lingual Machine Translation Approach: Inter-lingual MT approach intends to translate source language text to that of more than one language. Translation is from source language to an intermediate form called inter-lingual and then from inter-lingual to target language. Inter-lingual machine translation is one instance of rule-based machine-translation approaches. In this approach, the source language, i.e. the text to be translated, is transformed into an inter-lingual language, i.e. a language neutral representation. The target language is then generated out of the inter-lingual. One of the major advantages of this system is that the inter-lingual becomes more valuable as the amount of target languages it can be turned into increases. The inter-lingua approach is clearly most attractive for multilingual systems.

Transfer-based Machine Translation Approach: Transfer-based machine translation is similar to inter-lingual machine translation that it creates a translation from an intermediate representation that relate the meaning of the original sentence. Unlike inter-lingual MT, it depends partially on the language pair involved in the translation. On the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: Analysis, Transfer and Generation. In the first stage, the SL parser is used to produce the syntactic representation of a SL sentence. In the next stage, the result of the first stage is converted into

equivalent TL-oriented representations. In the final step of this translation approach, a TL morphological analyzer is used to generate the final TL texts. It is possible with this translation approach to obtain fairly high quality translations, with accuracy in the region of 90%. Three types of dictionaries are required: SL dictionaries, TL dictionaries and a bilingual transfer dictionaries.

2.5.2 Corpus-based Machine Translation Approach

Corpus based machine translation also called data driven machine translation is an alternative approach for machine translation to overcome the problem of knowledge acquisition problem of rule based machine translation [3]. Corpus Based Machine Translation (CBMT) uses, a bilingual parallel corpus to obtain knowledge for new incoming translation. This approach uses a large amount of raw data in the form of parallel corpora. This raw data contains text and their translations. These corpora are used for acquiring translation knowledge. Corpus based approach is further classified into the following two sub approaches [3]. Statistical Machine Translation approach and Example-based Machine Translation Approach.

Statistical Machine Translation Approach: SMT is generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The initial model of SMT, based on Bayes Theorem, proposed by Brown. Takes the view that every sentence in one language is a possible translation of any sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system.

The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by p(e|f), which is the Probability of translating a sentence f in the SL F (for example, English) to a sentence e in the TL E (for example, Ibo).

The problem of modeling the probability distribution p(e|f) has been approached in a number of ways.

One common approach is to apply Bayes theorem. That is, if p(f|e) and p(e)indicate translation model and language model, respectively, then the probability distribution $p(e|f) \propto p(f|e)p(e)$. The translation model p(f|e) is the probability that the source sentence is the translation of the target sentence or the way sentences in E get converted to sentences in F. The language model p(e) is the probability of seeing that TL string or the kind of sentences that are likely in the language E. This decomposition is attractive as it splits the problem into two sub problems. Finding the best translation is done by picking up the one that gives the highest probability:

$$e = \frac{argmax}{e \in e}\ p(e|f) = \frac{argmax}{e \in e}\ p(f|e)p(e)$$

SMT depends on a language model, a translation model and a decoding algorithm. The translation model ensures that the machine translation system produces target hypothesis corresponding to the source sentence. The language model ensures the grammatically correct output.

General architecture of SMT: The general architecture of SMT is shown in figure 2.2 [1]. The system source language text for translation. Language model, translation model and decoder attempts to process the source text and finally translated to target language text.

The translation model assigns a probability that a given source language sentence generates target language sentence. The training corpus for the translation model is a sentence-aligned parallel corpus of the languages.

Language model: tries to ensure that words come in the right order including some concept of grammar. The language model can be calculated with a statistical grammar or an N-gram language model. N-gram model was used for the purpose of the study. N-gram corpus is computed from monolingual corpus. The probabilities obtained from the N-gram model could be unigram, bigram, trigram or higher order N-grams.

Let's consider the following Afaan Oromo text:

Caalaan daabboo nyaatee

Caalaan shayee dhuge

Caalaan mana kitaabaa deeme

Caaltuun shayee dhugde

Alamuun buna dhuge

Unigram Probability can be calculated as follows: $P(O1) = \frac{Count(O1)}{Total\ Word}$

$$P(Caalaan) = \frac{count(Caalaan)}{Total\ word} = \frac{3}{16} = 0.19$$

Where 3 indicate appearance of the word Caalaan on the above example and 16 total number of words in the example above and 0.19 is the unigram probability of the word Caalaan.

Bigram probability can be computed as follow:

$$P(O2|O1) = \frac{Count(O1O2)}{count(O1)}$$

$$P(daabboo|calaan) = count(caalaan\ daabboo)/\ count(caalaan) = 1/3 = 0.33$$

Where 1 represent the word caalaan and daabboo appear together, and 3 represents the appearance of the word calaan in the example above and 0.33 the probability of the word caalaan and daabboo appears together based on the given text.

The trigram probability calculated as follow:

$$P(O3|O1O2) = \frac{count(O1O2O3)}{count(O1O2)}$$

$$P(nyaate|calaan\ daboo) = \frac{count(caalaan\ daabboo\ nyaate)}{count(caalaan\ daabboo)}$$

$$P(nyaate|caalaan\ daabboo) = \frac{1}{1} = 1$$

Where 1 indicate the appearance of the text caalaan daabboo nyaate, the appearance of the text caalaan daabboo and the result 1 indicate maximum probability. Based on the above information the language model compute the probability of the N-grams. The language model models the target language, that is, if the source text is English and the target text is Afaan Oromo, it models Afaan Oromo text.

For the corpus with simple sentences, the N-gram model performs well with the unigram, bigram and trigram models since the words in the sentence are not that long. But a problem exists if the sentences are too long and the solution would be smoothing which is avoiding zero probability. Which means by avoiding zero probability is no matter how long the decimal gets, it shouldn't be approximated to zero. Based on this method language model calculate the probabilities of N-grams which is used by decoder.

Translation model: assigns a probability that a given source language sentence generates target language sentence. As mentioned above, for a given source and target sentences **E** and O, it is the

way sentences in **E** get converted to sentences in **O** which is denoted by $P(O|E)$ calculated as follows:

$$P(O|E) = \frac{count(O,E)}{count(E)}$$

The above equation maybe difficult to achieve, if the sentences are too long. In order to overcome this problem the sentence is decomposed into smaller chunks, as in language modeling.

$$P(E|0) = \sum_{X} P(X,E|O)$$

The variable X represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be words or phrases. In word-based translation, the fundamental unit of translation is a word. Phrase-based translations, most commonly used, translates whole sequences of words, where the lengths may differ in which blocks are not linguistic phrases but, phrases found using statistical methods from corpus. The alignment probability $P(a,o|e)$ defined as follow:

$$P(a,o|e) = \prod_{j=1}^{n} y(aj|ei)$$

Where $y(aj|ei)$ is the translation probability and it is calculated by counting:

$$y(aj|ei) = \frac{count(aj,ei)}{count(ei)}$$

By following the above methods the language model and translation model calculate the probabilities distribution which is used by the decoder for translation.

Decoding: is a search for the shortest path in an implicit graph [1]. A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability, which is, $\frac{argmax}{o}\big(p(e|o) * p(o)\big)$ which indicate that

taking English text as an input and displays Afaan Oromo text as output. $\frac{argmax}{e}(po|e) * p(e))$ Which indicates taking Afaan Oromo text as an input and displays English text as output. By following the above procedures decoder perform the translations of the input text for both languages.

Finally decoder produces the best translation of the source language text according to the product of the translation and the language model. Finding the sentence that maximizes the translation and the language model probabilities is a search problem. A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability.

Source and target Text: in machine translation process for example if the translation performed from English text to Afaan Oromo text, English text is source text and Afaan Oromo is target text

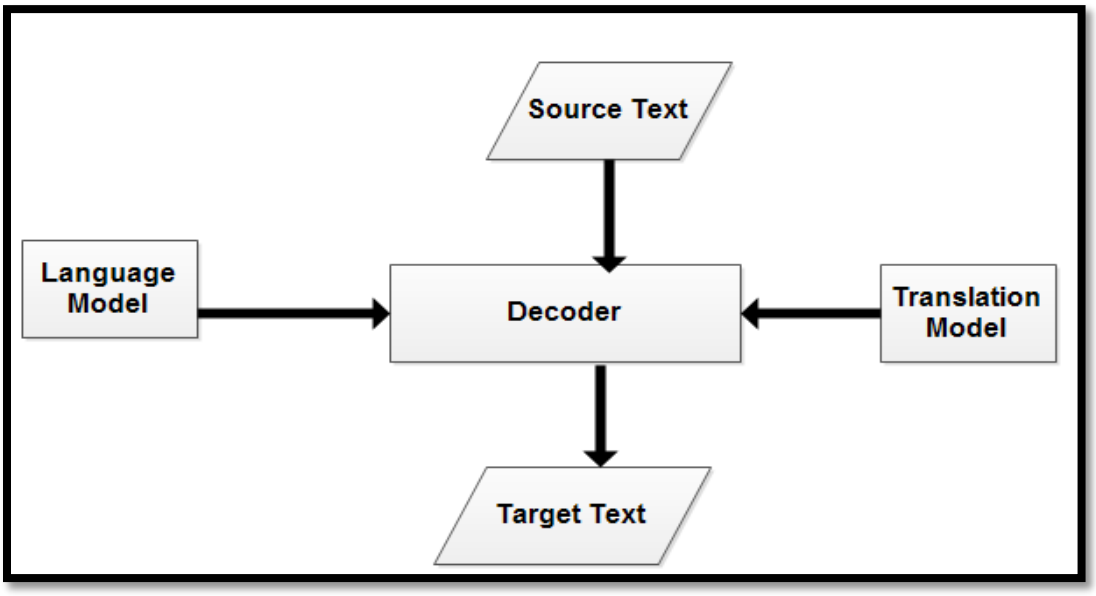

Fig 2.2: General architecture of SMT

How statistical machine translation work?: As indecated in chapter one machine translation has different approach, for this study we used statistical machine translaton approach because,

economically wise and also recommended by different researchers. Statistical machine translation is an approach that tries to generate translations using statistical methods based on bilingual text corpora. Statistical machine translation has three components [1]. Translation model, language model and decoder. The figure below shows the components of the approach:

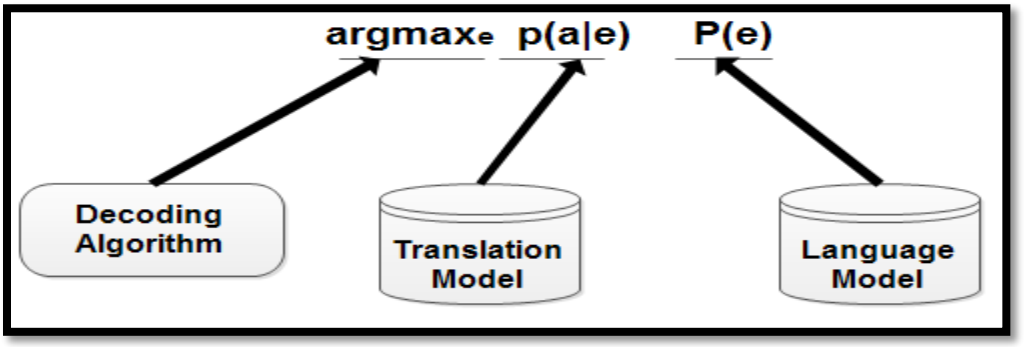

Figure 2.3: components of statistical machine translation

If we want to translate a sentence (a) in the source language (O) to a sentence (e) in the target language (E), the noisy channel model describes the process in the following ways: For example the translated sentence (a) must first considered in language (E) as some sentence (e**)** During communication (e) was corrupted by the channel to (a). Now, assume that each sentence in (E) is a translation of (a) with some probability, and the sentence that we choose as the translation (X) is the one that has the highest probability. $X = \frac{argmax}{e} P(e|a)$ Where $P(e|a)$ depends on one language model (types of the sentences found in language E) and second translation model (the way sentence E converted to sentence in A).

Derivation of Baye"s rule: $P(e|a) = P(a|e) * P(e)/P(a)$ where a is source text and e is target text $\frac{argmax}{e} P(e|a) = \frac{argmax}{e} P(a|e) * P(e)/P(a)$ by combining the questions we gate

$X = \frac{argmax}{e} P(a|e) * P(e)$ Which is used by the decoder for translation process.

Challenges of statistical machine translation: Some challenges in SMT includes [3], Sentence alignment; in parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence alignment can be performed through different alignment algorithm. Statistical Anomalies: Real-world training sets may override translations of,

say, proper nouns. An example would be that "I took the train to Berlin" gets mis-translated as "I took the train to Paris" due to an abundance of "train to Paris" in the training set. Corpus creation can be costly for users with limited resources, the results are unexpected, Statistical machine translation does not work well between languages that have significantly different word orders (e.g. Japanese and European languages).

Evaluation procedure of SMT: SMT evaluated by using human evaluation and automatic evaluation method. BLUE score is one of automatic evaluation metric in order to evaluate the performance of SMT. The algorithm perform the computations as follows [3]:

BLEU computation is based on two elements: The first one is N-gram Precision; the N-grams output by the machine translation system, what percentage appear in a reference sentence? And the second is Brevity Penalty; the brevity penalty puts a penalty on sentences that are shorter than the reference, preventing these short sentences from receiving an unnecessarily high score.

First it defines $e = e1 \dots \dots \dots en$ as an arbitrary N-gram of length-n. Then it defines a function $occour(E, e)$ number of times that e occurs in sentence E. Then it defines two function an N-gram count function that counts the number of N-grams of length-n in the system output E: $Count(b, n) = \sum_{e \in \{e; |e|=n\}} occur(b, e)$ and N-gram match $match - n(E, b, n)$ counts the number of times that a particular N-gram occurs in both the system output and reference. $match(E, b, n) = \sum_{e \in \{e; |e|=n\}} \min(occur(E, e), occur(b, e))$. Then, given a corpus to system outputs c and references d, it accumulate the counts and matches over each sentence in the corpus.

$$count(c, n) = \sum_{b \in d} count(b, n)$$

$$match(d, c, n) = \sum_{\{E,b\} \in \{c,d\}} match(E, b, n)$$

Then calculate the N-gram precision for the corpus as the number of matches divided by the number of N-grams output:

$$prec(c, d, n) = \frac{match(c, d, n)}{count(c, n)}$$

The brevity penalty is designed to penalize system outputs that are shorter than the reference, and is multiplied with the N-gram precision terms of the BLEU score, so a lower value for the brevity penalty indicates that the score will be penalized more and calculated with the following equation:

$$brev(c,d) = \begin{cases} 1 \\ e^{1-\frac{count(d,1)}{count(c,1)}} \ if\ count(c,1) > count(d,1) \end{cases}$$

$if(count(c,1) > count\ (d,1))$, no penalty will be imposed when the output is longer than the reference, and the penalty reduces the score to zero as the length ratio reduces to zero.

Finally, combining all of these together, it takes the geometric mean of the N-gram precisions up to a certain length of n (almost always 4) and multiply it with the brevity penalty to get BLUE score:

$$BLUE(d,c) = brev(d,c) * \exp(\sum_{n=1}^{4} log\ prec(d,c,n))$$

Example-based Machine Translation (EBMT), Approach EBMT is characterized by its uses of bilingual corpus with parallel texts as its main knowledge, in which translation by analogy is the main idea [3]. An EBMT system has given a set of sentences in the source language and corresponding translations of each sentence in the target language with point to point mapping. These examples are used to translate similar types of sentences of source language to the target language. There are four tasks in EBMT: example acquisition, example base and management, example application and synthesis.

The principle of translation by analogy is encoded to example-based machine translation through the example translations that are used to train the system. Challenges of EBMT approach [3]; EBMT is an attractive approach to translation because it avoids the need for manually derived rules. However, it requires analysis and generation modules to produce the dependency trees needed for the examples database and for analyzing the sentence. Another problem with EBMT is computational efficiency, especially for large databases, although parallel computation techniques can be applied.

2.5.3 Hybrid Machine Translation Approach

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach [3]. Which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT

sectors use this hybrid based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system.

2.6 Sentence alignment

Alignment is matching of the two texts, typically into small logical units such as sentences or phrases, such that the *n*th segment of the first text and the *n*th segment of the second are mutual translations [17]. Sentence alignment is the problem of making explicit the relations that exist between the sentences of two texts that are known to be mutual translations. Sentence alignment is the problem of, given a parallel text, finding a bipartite graph matching minimal groups of sentences in one language to their translated counterparts [18].

2.6.1 Impact of sentence alignment on SMT

The quality of a statistical machine translation (SMT) system is heavily dependent upon the amount of parallel sentences used in training [19]. For any statistical machine translation system, the size of the parallel corpus used for training is a major factor in its performance.

Sentence-aligned parallel bilingual corpus have proved very useful for applying to machine translation, but they usually do not originate in sentence aligned form. This makes the task of aligning such a corpus of considerable interest, and a number of methods have been developed to solve this problem [20]. Based on the above concepts the alignment of parallel corpus affect the performance of the machine translation especially on SMT.

2.6.2. Tools used for sentence alignment

Literature shows that there are different tools developed for aligning corpus for different purpose of text processing [21] [22] [23]. The following are some common tool:

2.6.2.1. Giza++

GIZA++ is part of the SMT toolkit EGYPT which was developed by the SMT team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University [21].

Giza++ is an extension of an older program, Giza. This tool performs statistical alignment, implementing several Hidden Markov Models and advanced techniques allow to improve alignment results. GIZA++ is part of the statistical machine translation toolkit used to train IBM Model 1 to Model 5 and the Hidden Markov Model [22].

IBM Model 1

IBM Model 1 is the simplest model among the models that the IBM team proposed to estimate lexical translation using the probability equation below:

$p(e,a|f) = \epsilon \frac{\epsilon}{(lf+1)le} \prod_{j=1}^{le} t(ej|fa(j))$ Where

ϵ - Is normalized constant

$f = (f1, \ldots flf)$ – Foreign sentence of length of l_f

$e = (e1, \ldots ele)$ - English sentence of length l_e

With an alignment of each English word e_j to a foreign word f_i according to the alignment function $a: j \rightarrow i$

IBM Model 1 is weak in terms of conducting reordering or adding and dropping words. In most cases, words that follow each other in one language would have a different order after translation, but IBM Model 1 treats all kinds of reordering as equally possible. Another problem while aligning is the fertility (the notion that input words would produce a specific number of output words after translation), as shown in the figure 2.4.

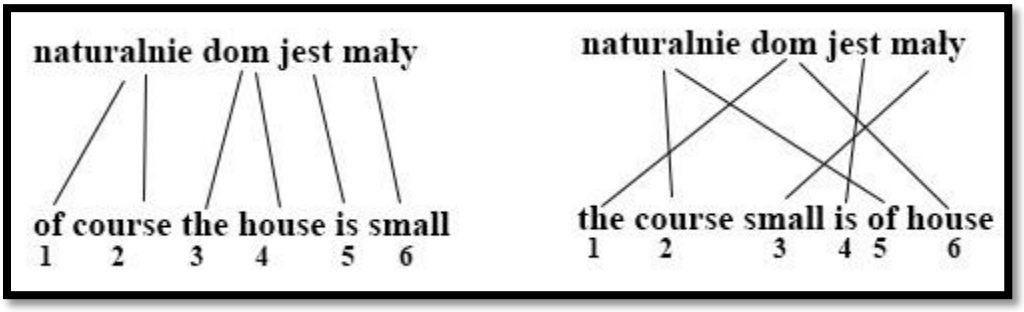

Figure 2.4: Alignment probability using IBM model 1

IBM Model 2

The limitation of Model 1 doesn't consider where the words appear in either of the strings. Therefore, Model 2 builds on top of Model 1 to reorder the words in the target sentence.

Translating foreign word at position i to English word at position j: $a(i|j, le, lf)$ by combining with IBM model 1 the probability equation becomes:

$$p(e, a|f) = \in \prod_{j=1}^{le} t(ej|fa(j))a(a(j)|j, le, lf)$$

The translation done by IBM Model 2 can be presented as a process divided into two steps (lexical translation and alignment).

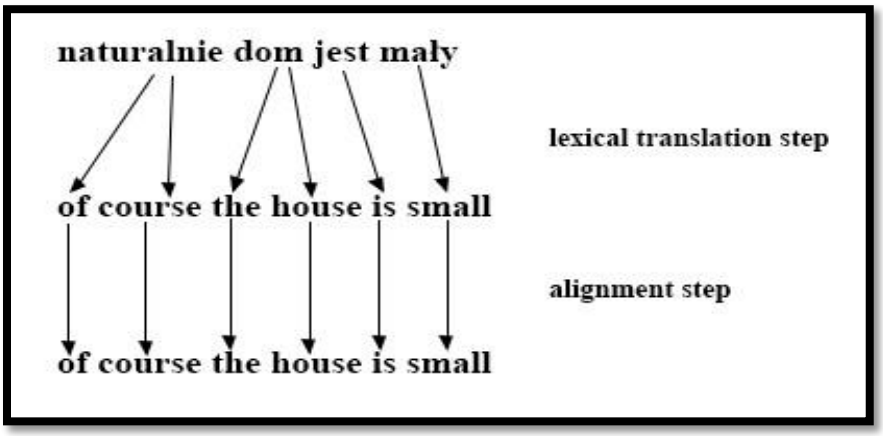

Figure 2.5: Lexical translation and alignment probability using IBM model 2

In this equation, the alignment function maps each output word (j)to a foreign input position .

IBM Model 3

A single words in the source language may not map to exactly one word in the target language. Model 3 adds the fertility probability n(sj) which is equal to the likelihood of each source word translated to one word, two words, three words, and so on, on top of Model 2 parameters Modelled by distribution $n(\emptyset|f)$. The number of inserted words depends on sentence length. This is why the NULL token insertion is modeled as an additional step to the fertility step. It increases the IBM Model 3 translation process to four steps as shown in the figure 2.6:

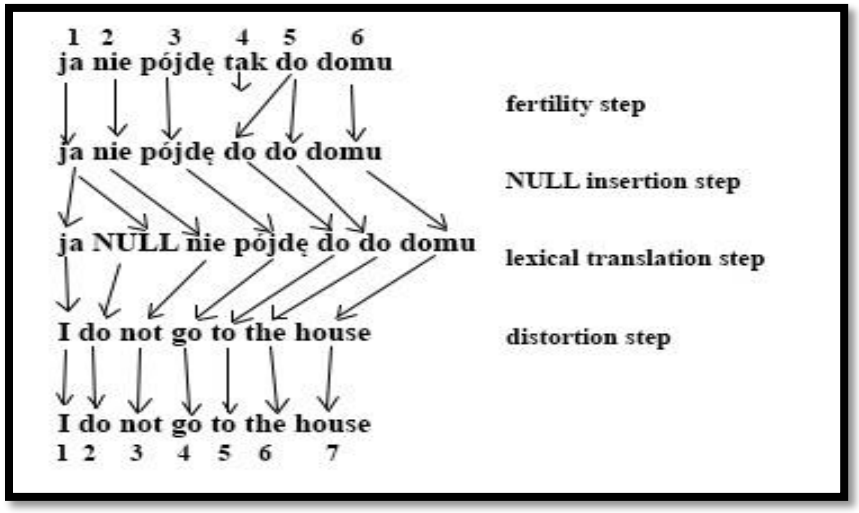

Figure 2.6: Alignment probability using 4 steps IBM model 3

IBM Model 4

The set of distortion probabilities for each source and target position (i.e., the probability of a word in the source sentence change its position in the target sentence). As opposed to Model 2 which does absolute reordering, model 4 does relative reordering.

IBM Model 5

Model 5 removes the deficiencies of the previous models [1-4]. For example, Model 4 can stack several words on top of one another. It can also place words before the first position or beyond the last position in the target string. Therefore, Model 5 fixes deficiencies like this one that the previous models have not handled.

2.6.2.2 Hunalign

Hunalign is a sentence-level aligner built on top of [Gale and Church]'s algorithm written in C++. When provided with a dictionary, hunalign uses its information to help in the alignment process, despite being able to work without one [21]. Hunalign implements an alignment algorithm based on both sentence length and lexical similarity. In general similar with Moore's algorithm.

The main difference is that hunalign uses a simple word by word dictionary based replacement instead of IBM model 1. On one hand this results in significant speed gains. More importantly, however, it provides flexible dependence on the dictionary, which can be pre-specified (if one is available) or learned empirically from the data itself. In case a dictionary is not available, an initial pass is made, based only on sentence length similarity, after which the dictionary is estimated from

this initial alignment and a second pass, this time with the dictionary is made. Consumption is its weak spot; in reality it cannot handle parallel corpora larger than 20 thousand sentences [23].

2.6.2.3. Gale and Church Algorithm

The Gale and Church algorithm is based on character based sentence length correlations, i.e. the algorithm tries to match sentences of similar length and merges sentences, if necessary, based on the number of words in the sentences [23]. The alignment model proposed by Gale and Church (1993) makes use of the fact that longer/shorter sentences in one language tend aligned into longer/shorter sentences in the other. A probabilistic score is assigned to each proposed sentence pair, based on the sentence length ratio of the two sentences (in characters) and the variance of this ratio.

This probabilistic score is then used in the dynamic programming framework to get the maximum likelihood alignment of sentences.

2.6.2.4. Gargantua

Gargantua aims to improve on the alignment algorithm by Moore (2002) by replacing the second pass of Moore's algorithm with a two-step clustering approach [23]. As in Moore's algorithm, the first pass is based on sentence-length statistics and used to train an IBM model. The second pass, which uses the lexical model from the first pass, consists of two steps. In a first step, a sequence of 1-to-1 alignments is obtained through dynamic programming. In a second step, these are merged with unaligned sentences to build 1-to-many and many-to-1 alignments.

2.6.2.5 Bleualign

Bleualign uses an automatic translation of the source text as an intermediary between the source text and the target text [23]. A first alignment is computed between the translated source text and the target text by measuring surface similarity between all sentence pairs, using a variant of BLEU, then finding a path of 1-to-1 alignments that maximizes the total score through dynamic programming. In a second pass, further 1-to-1, many-to-1 and 1-to-many alignments are added through various heuristics, using the alignments of the first pass as anchors. Bleualign does not build its own translation model for the translation of the source text, but requires an external MT system.

2.6.2.6 Bilingual Sentence Aligner

The Bilingual Sentence Aligner [23] combines a sentence length based method with a word correspondence based method. While sentence alignment based on sentence-length is relatively

fast, lexical methods are generally more accurate but slower. Moore's hybrid approach aims at realizing an accurate and computationally efficient sentence alignment model that is not dependent on any additional linguistic resources or knowledge.

The aligner implements a two-stage approach. First the corpus is aligned based on sentence length. The sentence pairs that are assigned the highest probability of alignment are then used as training data for the next stage. In this second stage, a lexical model is trained, which is a modified version of IBM model 1. The final alignment model for the corpus combines the initial alignment model with IBM model 1 [23].

2.6.2.7 Anymalign

Anymalign is a multilingual sub-sentential aligner. It can extract phrase equivalences from parallel corpora. In order to extract phrases it uses punctuation mark such as comma and hyphen as delimiters or end of line to align phrases of both target and source language. Its main advantage over other similar tools is that it can align any number of languages simultaneously.

Some characteristics of Anymalign are the following [24]: It is truly multilingual; any number of languages can be aligned simultaneously. Anymalign also fast; Quality of results is not a matter of time, however coverage is. The longer Anymalign runs, the more results. The program can be stopped at any time. It is easy to use (a single command should suffice for most purposes), easy to parallelize (just run the very same command on several machines their results can be merged with a single command) and easy to integrate (simple one-file input and output formats). There is no intermediary step. Portable: written in the Python programming language, available for most systems.

From the alignment tools mentioned above we used GIZA++, Anymalign and hunalign for word level, phrase level and sentence level alignment respectively because, these tools goes with our objective and they are current tools used in SMT research area.

2.7 Related works

Different research conducted on machine translation based on different approach and methodology for both local and foreign languages. The following are some researches that relate with machine translation and related with our study:

2.7.1 English-Amharic statistical machine translation

This experiments was conducted by Mulu and Besacier [13]. The demand for translation has become more in recent years due to increase in the exchange of information between various regions using different regional languages. Accessibility to web document in other languages, for instance, has been a concern for information Professionals.

The author use corpus associates probabilities with translations empirically by counting co-occurrences in the data. Estimates of probabilities get more accurate as the size of the data increases. Most translation systems use parallel corpus for training data from constitutions called Hansard Corpus. Similarly, the English-Amharic parallel corpus from parliamentary documents that exist online including those collected manually are used for the preliminary experiment on EASMT.

The pre-process convert the corpus from PDF to RTF then to Unicode text. The number of successfully converted corpora from the total 632 is 115. The low numbers is due to some of the oldest Gazeta, which are saved as jpg image formats. Aligning into Amharic and English is already done as both are incorporated on the same page. The most challenging task was converting from the RTF to Unicode text file. This is because each corpus can have at least 8 different Amharic fonts. If a word contains more than two fonts during conversion, then the converter automatically converts the word using the first encountered font. The words with other fonts contain weird characters after the automatic conversion of the full document is complete wrongly converted words are manually corrected.

Trimming has been performed by removing any part of the corpus except the text that contains the full content of the proclamation. After automatically trimming the corpora, the process of splitting each paragraph into sentences using sentence endings is performed. The Amharic sentence endings and punctuations have been converted to English to make it easy to apply similar pre-processing tools used for English. The converted Amharic punctuations include the Ethiopic comma (፣), colon (፡), semi-colon (፤) and full stop (።) to their English counterparts (‘,’, ‘:’, ‘;’, ‘.’) respectively.

The alignment at the sentence level has been done using a sentence aligner called Hunalign. Hunalign aligns bilingual text at sentence level using sentence-length information. A small

English-Amharic bilingual dictionary, which is adopted from word lists sized 8,212 have been used. The aligner was able to align 19,115 Amharic sentences and 25,730 English sentences.

The aligner was able to align 18,434 English Sentence to Amharic sentences in 0-1, 1-1, or 1-2. Those sentences that do not have matching translations (0-1 or 1-0) have been dropped. Those sentence pairs with more than 200 tokens in length have been dropped as well in order to get a better performance of the decoder.

Out of the total collected data, 90% or 16,432 randomly selected sentence pairs have been used for training while the remaining 10% or 2,000 sentence pairs are used for tuning and testing. Thus, the preliminary experiment is developed using a total of 18,432 English-Amharic bilingual parallel and 254,649 monolingual corpora. The monolingual corpus is used for the Language Modeling (LM). The LM corpus contains those data related to parliamentary documents that are not included in the bilingual corpora and news items collected from the Ethiopian News Agency.

The other resources used are statistical analysis toolkits such as word alignment, language modeling (LM), translation modeling, and evaluation. The toolkit used to build the language model is SRILM. Whereas to build the statistical translation model, an open source Giza++ toolkit as well as some scripts that are with the Moses suite are used. The BLEU metric is used to evaluate the performance of the test result. All mentioned toolkits are integrated with MOSES.

The author use English-Amharic parallel Training Set as translation examples and tested using the English Source Text as new sentences that gives an output Target Text of translated Amharic sentences. Based BLEU score result indicates that 35.32% translation BLEU scores have been achieved.

The author suggest at the end, more experimentation and research is required to further improve the translation accuracy of the EASMT.

2.7.2 Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus

The thesis was conducted by Eleni Teshome for partial fulfilment of MSC Addis Ababa University, 2013 [1].

The objective of this study was to design and develop a bi-directional English-Amharic machine translation system using constrained corpus.

She used Sample text corpora from relevant data sources with parallel text. For the simple sentences, 1020 sentences were manually prepared and for the complex sentences, 1951 were collected. That is, 414 from the Public Procurement Directive and 1537 sentences from the Bible.

For the study she used, Statistical machine translation approach. Statistical machine translation is an approach that tries to generate translations using statistical methods based on bilingual text corpora. Researchers discovered that as the size of the corpora increases, the accuracy of the translation improves.

Two methodologies are used to test the system. The first one is BLEU Score and the second methodology is preparing a questionnaire manually. BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.

The second methodology she followed was using a manual technique. The questionnaire consisted of all the testing dataset and the translation acquired. It has an evaluation method on the scale of 1 to 5 in which if a candidate gives 5, it means that the translation was perfect and if 1 is given, it means it has a very poor translation. Thirteen candidates were selected to fill the questionnaire. Two questionnaires were prepared for the test set i.e. from English to Amharic and from Amharic to English. The questionnaires developed were different because the result obtained from Amharic to English and from English to Amharic were not similar.

The author seen results from two perspectives, one from the accuracy point of view and the other from the time it takes to translate a particular sentence. From the experiments taken, the following findings were presented.

Similar to the experiment, the corpora was named corpus I and corpus II. For each experiment taken, the result was recorded. For corpus I, the following result was obtained for the training set as well as the test set.

The result recorded from the first methodology (BLEU Score) was 82.22% for the English-Amharic translation and 90.59% for the Amharic-English translation. For the second methodology, the result recorded was 91% for the English-Amharic translation and 97% for the Amharic to English. The time it took for each translation to take place was recorded and for the English-Amharic translation, the highest time it took was 17 microseconds. For the Amharic-English translation as well, the maximum time recorded was 0.009.

The following are the results acquired from corpus II for the merged training and test sets. Results were obtained from the complex sentence that was taken as a sample text from Corpus II. As mentioned above, forty sentences were taken as a sample, that is, both from the directive and the Bible. For the first methodology, the accuracy of the translation from English to Amharic was 73.38%. And the translation from Amharic to English was 84.12% effective. The time it took for each result to be produced is an average of 4.987 seconds.

Finally the author recommend that, further researches in machine translation on Amharic to other languages, in Ethiopia such as Tigrigna, Afaan Oromo or so could be performed while preparing a large corpus.

2.7.3 English-Afaan Oromo machine translation: An experiment using statistical approach

The study was conducted by Sisay Adugna in 2009 with the objectives of to apply existing SMT systems on English – Afaan Oromo language pair by using available parallel corpus and to identify the challenges that need a solution regarding the language pair [8].

Lack of utilization or accessibility of online collection for information need of Afaan Oromo speakers is considered as the main problem by the author.

The researcher found digitally available Afaan Oromo versions of some chapters of the Bible and some spiritual manuscripts for which the English counterparts were explored on the web. Another typical bilingual text is the United Nation's Declaration of Human Rights, which is available in many of the world's languages, including Afaan Oromo.

The researcher used parallel documents from different domains including spiritual documents and legal documents. 20,000 bilingual sentences and 62, 300 monolingual sentences were used for training and testing purpose. The data is organized into training and testing data in the proportion of 9:1 (90% for training and 10% for testing).

The training of the system was done is 32 bit Linux machine as an operating system platform.
He used SRILM toolkit and GIZA++ for language modeling and word alignment respectively, the Moses decoder was used for decoding purpose. The documents were preprocessed by using different scripts which are customized to handle some special behaviors of Afaan Oromo such as apostrophe. Sentence aligning, tokenization, lowercasing and truncating long sentences that take the alignment to be out of optimality were also done by those scripts.
In order to evaluate the performance the system he use BLUE score which is automatic method used to evaluate the system.

The author performs different experiments by varying number of N-grams, the highest BLEU score of 43.96% is observed for 1- gram scoring. However, the N-gram score sharply drops as N increases, i.e., the N-gram score for values of N equals 1, 2, 3, 4, 5, 6, 7, 8 and 9 is observed to be 43.96%, 21.57%, 14.42%, 10.72%, 8.04%, 5.52%, 3.76%, 2.23% and 1.30% respectively and finally the result of BLUE score also differ based on the domain of the data collected.
At the end the researcher recommend that, using better alignment tool for the corpus used for training and testing resulted to better performance and some Afaan Oromo words in the corpus were considered by the system as different words due to spelling errors. Therefore, he strongly recommend the development of spell checker for Afaan Oromo that will help facilitation of the document preparation.

2.7.4 Bidirectional English-Afaan Oromo Machine Translation Using Hybrid Approach

The research was conducted by Jabesa Daba in 2013, with the objective of developing a bi-directional English-Afaan Oromo machine translation system using hybrid approach [9].

The author uses the following source of corpus for study; Holy Bible, the Constitution of FDRE, the Criminal code of FDRE, International conventions, Megeleta Oromia and a bulletin from Oromia health bureau. He collect corpus from different web sites and different offices. After he prepare corpus the author performs Tokenization, True-casing and cleaning to make ready for experiment. He uses a total of 3000 corpus from this he use 90% for training and remaining 10 percent for testing for both approaches.
The result recorded from the BLEU score methodology shows 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation by using statistical approach. The

result recorded from the BLEU score methodology shows that 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation by using rule based approach.

From the result of both the experiments the Author conclude that, result recorded from a BLEU score shows that the hybrid approach is better than the statistical approach for English-Afaan Oromo language pair.

The author recommend that, the rules which are developed and used in the system are only used for syntax reordering. Therefore, additional results can be accomplished by further exploring the rules especially by developing morphological rules.

2.7.5 Intelligent hybrid man-machine translation Evaluation

The research was conducted by Ibrahim Ahmed and Ibrahim Saleh Sabek (2014). The objective of introducing a hybrid machine translation evaluation approach that addresses the drawbacks of automatic metrics and benefits from human assessments [25].

The metrics they used to evaluate the output of the proposed machine translation evaluation approach can be divided into two main categories: accuracy metrics and time metrics.

The accuracy metrics include two different metrics: correlation with human judgment and cumulative distribution function of correlation.

The proposed approach is implemented using Java 6.0 as the programming platform. For word alignment, they trained IBM Model-4 using GIZA++. Using the monolingual corpus, they trained many English language models with different N-grams. For the online operation, their approach uses extracted phrases corresponding to both tokenized monolingual corpus and word alignments from the offline phase.

They used two types of datasets: Europarl multilingual corpus as a large dataset, and WMT corpora as small datasets.

The majority of the training data in the two datasets was drawn from the Europarl corpus. Additional training data was taken from the News Commentary corpus the domain is general politics, economics and science. Evaluation results are divided into two parts. The first part evaluates the performance of the proposed approach using the linguistic and data-driven features and without including human assessments.

The second part studies the effect of incorporating human assessments on the overall accuracy.

In this part, only WMT datasets are used because no human assessments are available.

The results of the study show that the proposed approach has the best performance under the two datasets. The proposed approach provides an enhancement of at least 13.92% in accuracy over the best state-of-the-art techniques for the WMT 2007 datasets and at least 3.85% for the WMT 2013 datasets. All techniques perform worse in WMT 2013 dataset due to the high complexity of training and testing sentences in this dataset.

2.7.6 Chinese-English Statistical Machine Translation by Parsing

The research conducted by Yue Zhang Mansfield in (2006), with the aim to study large-scale Chinese-English SMT using a syntactic tree-based model [26].

GenPar is the major software framework used for the experiments. It provides an implementation of the generalized parsing algorithm. GenPar is written in C++. It was developed during the 2005 Johns Hopkins Language and Speech Processing Workshop.

To train a statistical GMTG, a corresponding bilingual treebank is needed. This bilingual treebank is not directly available, and can be derived from mono-lingual syntax trees using the hierarchical alignment algorithm, which requires word-to-word translation probabilities. After maximum likelihood training using the bilingual treebank, EM based training can be conducted to optimize the probabilistic grammar.

The Bikel parser is used as mono-lingual parser that supports different types of statistical parsing models. It is used in the experiments to produce English and Chinese mono-lingual grammars. This parser is implemented in Java. The Bikel parser requires the part-of-speech information for each word in the input sentence. Therefore, before using the parser, POS tagging (Jurafsky and Martin, 2000) is required, which assigns POS tags to the input words. GIZA++ used as word alignment tool.

Hong Kong Parallel Text corpus is used for the bilingual training. It is produced by the Linguistic Data Consortium (LDC). The corpus contains parallel articles in Chinese and English, together with specifications of the sentence alignment in each pair of articles. There are three collections in the corpus, containing 59 million Chinese words and 49 million English words. Only the News collection is used for the study, which contains 27 million Chinese words and 15 million English words (in 2,681 thousand Chinese sentences and 2,952 thousand English sentences, respectively). The Hong Kong Parallel Text corpus is used for part of the testing data. Parallel sentences in the corpus that are not used for training are extracted for testing. The Multiple-Translation Chinese

(MTC) Part 3 corpus is used for another part of the testing data. The corpus contains 100 Chinese articles, each having four independent English translations. The corpus is sentence-aligned, and there are 935 Chinese sentences in total.

In order to evaluate the translation performance the author used BLUE score, NIST score and F-measure.

Training set size (sentence pairs)	FMS-1	FMS-2	BLEU	NIST
1k	0.2477	0.1457	0.0155	0.5382
5k	0.3196	0.1711	0.0394	2.0645
10k	0.3310	0.1720	0.0437	2.8120
50k	0.3687	0.1841	0.0604	4.0294
100k	0.3742	0.1850	0.0605	4.1228

CHAPTER THREE

Overview of Afaan Oromo and English language

3.1 Overview of Afaan Oromo language

Afaan Oromo is one of the languages of the Low land East Cushitic within the Cushitic family of the Afro-Asiatic Phylum [27, 28]. It is also one of the major languages spoken in Ethiopia. According to Gene [29] and Hamid [30], 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 [31]. 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 [31]. In spite of its usage, as a vernacular, the language is widely spoken in the Horn of Africa [31].

Like Amharic language, Afaan Oromo is rich in morphology. That is, the language in which significant information concerning syntactic units and relations is expressed at word-level. [32]. Latin based alphabet known as Qubee has been adopted and became an official script of Afaan Oromo since 1991 [33]. The language is widely used in Ethiopia and neighboring countries like Kenya and Somalia [34]. Currently, Afaan Oromo is an official language of Oromia Regional State and used as an instructional media for primary and junior secondary schools of the region. Even if the language is spoken by large number of the population, the number of literature works, newspapers, magazines, educational resources, official documents and religious writings written and published in this language are few in number.

3.2 English-Afaan Oromo Linguistic Relationship

English and Afaan Oromo have some structural differences as well as similarities. The syntactic and lexical relationships that are challenging during translation from English - Afaan Oromo and from Afaan-Oromo to English are discussed below.

3.2.1 Noun

Nouns in Afaan Oromo can vary to reflect number, gender, and case (subjective, objective or possessive). Number of noun can be indicated by using numerals and quantifiers.

Usually, number in countable nouns is indicated by numerals and quantifiers, for example, in mana shan (five house), shan (five) indicates the number of house and in mannen baay'ee (many house) baay'ee (many) is the quantifier to reflect plurality.

In addition to numerals and quantifiers, plural markers also indicate the number of noun. Unlike English that mainly has a single plural marker (suffixing's' to a noun), Afaan Oromo has many plural markers such as *-oota*, *-olii*, *-lee*,and -w*wan*. In Afaan Oromo, a plural noun is produced by suffixing these plural markers on the singular noun. Example Nama-Namoota, Laamii-Lamiilee, sa'a - saawwan etc.

3.2.2 Personal Pronouns

The personal pronouns as subjects and direct objects are listed below along with possessive markers. Like English, Afaan Oromo uses different forms of personal pronouns to indicate their role in the sentence. While "he" and "him" may refer to the same person, English uses "he" for subjects and "him" for objects and for Afaan Oromo "Isheen, Inni, Nuti" used for object. Afaan Oromo has several forms for all nouns, including pronouns forms [35]. Examples:

She likes him-----Isheen isa jaalatti

He likes her-----Inni ishee jaalata

We buy it-----Nuti isa binna

Do you hear me? -----Ati na dhageessa? Or more commonly Na dhageessaa?

3.2.3 Adjectives

Afaan Oromo adjectives can be male, female, or neutral. Masculine adjectives are used with masculine nouns, feminine adjectives modify feminine nouns, and neutral adjectives can be used with any noun. All non-neutral adjectives can be made masculine or feminine by attaching the appropriate suffix. Masculine suffixes for adjectives are: *-aa*, *-aawaa*, *-acha*, and *-eessa*. Feminine suffixes are: *-oo*, *-tuu*, *-ooftuu*, and *-eettii*. Standard morphology rules apply when attaching suffixes.

Example:

English meaning	Masculine	Feminine
Adorable	Jaallatamaa	Jaallatamtuu
Beautiful	Bareedaa	Bareedduu
Fast	si'aawaa	si'ooftuu

Sweet	mi'aawaa	mi'ooftuu
Fat	Furdaa	Furdoo
Small	xinnaa, xiqqaa	xinnoo, xiqqoo
Messy	Boosacha	Booseettii
Black	Gurraacha	Gurraattii
Poor	Hiyyeessa	Hiyyeettii
Skinny	Godeessa	Godeettii

Neutral adjectives (e.g., **adii** – "white") use the same form for both masculine and feminine nouns.

3.2.4 Afaan Oromo and English Sentence Structure

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, then the object and the verb next to the object. 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 [35].

3.2.5 Articles

English language uses two types of articles known as definite article (the) and indefinite article (a, an, some, any). In case of Afaan Oromo, there are no articles that are inserted before nouns unlike that of English rather the last vowel of the noun is dropped and suffixes (-icha, -ittii, -attii, -utti) are added to show definiteness instead of using definite articles as shown in the example below:

karaa ------- 'road', karicha---------- 'the road'

Nama -------'man', namicha/namticha------- 'the man'

Haroo------ 'lake', harittii ----------'the lake'

Qaalluu -----'priest', qaallicha----- 'the priest

Qallittii---------- 'the priest'

3.2.6 Punctuation Marks

Other than apostrophe, punctuation marks used in both Afaan Oromo and English languages are the same and used for the same purpose. Apostrophe mark (') in English shows possession, but in Afaan Oromo it is used in writing to represent a glitch sound known as hudhaa.

It plays an important role in Afaan Oromo reading and writing system. For example, it is used to write a word in which most of the time two vowels appear together like “ta'uu”.

3.2.7 Modifiers

Adjectives come after the nouns they modify. Adverbs that modify adjectives go before the adjective. Adverbs that modify verbs, adverbial clauses, and relative clauses tend to go at the beginning of the sentence before the subject [35].

Examples:

Your blue pen is in my room ----- biirii dooqee kee kutaa koo keessa jira

I don't know where it is ----- Eessa akka ta'e ani hin beeku

How far is the post office? ----- Hagam manni postaa fagaata?

What are you doing tonight? ----- Edana maal gotta?

3.2.8 Verb Groups for Conjugation

Most Afaan Oromo dictionaries list verbs in their infinitive from (e.g., beekuu - “to know”), and all infinitives end in *-uu*. The verb stem is this infinitive form with the final *–uu* dropped. The stem of beekuu is therefore *beek-*, and the verb is conjugated by adding suffixes to this stem (e.g beekti - “She knows”). Afaan Oromo verbs fall into one of four groups based on their stem ending [35].

Group 1: Regular Verbs: Most verbs in Afaan Oromo are “regular”, that is, they attach the regular person and number-based suffix to their stem without any changes to the stem or suffix. These are verbs with stems that do not end in: a double consonant, *ch*, a vowel, *y*, or *w*. the present future conjugations for deemuu are shown below as an example with suffixes.

Verbs that don't fall into one of the other three categories follow this pattern of conjugation.

Group 2: Double-consonant Ending Stems**:** If the verb stem ends in a double consonant, a slight modification of the regular verb conjugation must be made because, Afaan Oromo does not allow three consonants to occur in a row. For nuti, ati, isin, and isheen forms, an, I is added to the regular suffix. Other verbs that follow this pattern include: arguu, gadduu, rommuu, and gorsuu.

Group 3: *-chuu* Verbs**:** Many verb infinitives end with *-chuu*. For these verbs, the *ch* changes to *dh* in the ani form and to *t* for all other forms. Then the standard suffixes are applied. The example of nyaachuu is given below with stem changes and suffixes. Note that the *t* changes to *n* for the

nuti form. Other verbs in this group include: jirachuu, fudhachuu, argachuu, guddifachuu, barachuu, and gubachuu.

Group 4: Vowel-Ending Stems (Irregular Verbs). Infinitives that end with *-a'uu* , *-o'uu*, *-u'uu*, *-e'uu*, and *-i'uu* behave as regular verbs for ani, inni, and isaan froms. However, for the other forms, the stem and/or suffix will deviate from regular conjugations.

3.2.9 Comparatives

Comparative is the form of adjective or adverb used to compare two things. There is no direct translation of the English *-er* in Afaan Oromo. Most often, when distinguishing between two objects, as in "the longer one", the Afaan Oromo phrase would simply be "the long one" ("isa dheera") or "the very long one" ("isa baay'ee dheera"). Baay'ee, in addition to meaning "very", can also convey the sense of "more" when used with an adjective.

The adjective caalaa can be used to mean "better" or "more", though most often it functions as an adverb and comes immediately before the verb, as in "Isheen caalaa bareeddi" ("She is more beautiful"). Caalaa comes from the verb caaluu meaning "to be better". "Inni caala" thus means "its better". Some dialects may use daran instead of caalaa as a comparative adjective/adverb.

The preposition irra, meaning "on", can signify a comparison in a way that more literally means "relative to". For example, "Inni ishee irra gabaabaa dha" means

"He is shorter than she (her)" or "He, relative to her, is short". In many cases, caalaa can be added to irra for optional emphasis, as in "Finfinneen Maqalee irra (caalaa) bareeddi?" ("Is Finfinnee more beautiful than Mekele?"). Note that cities are treated as feminine. For "worse", gadhee ("bad") is most often used, as in "sun kanarra gadhee dha" ("that is worse than this").

For equating two things, as in "as good as" or "as <any adj.> as", akkuma can be used. "Chelsii akkuma Manchester gaarii dha" thus means "Chelsea is as good as Manchester". Akka can also be used to mean "like" or "similar to", as in "Chaaltuun akka Hawwiittuu barattuu dha" ("Chaltu is a student like Hawitu"). Additionally, hanga (haga in some dialects) means "as much as", as in "Bilisaan hanga Argaayaa beeka" ("Bilisa knows as much as Argaya").

Note that akka and akkuma come between the nouns being compared. When two things being compared are objects (e.g., "He likes this more than that"), irra comes after the first object. When one item is the subject and the other an object (e.g., "This is better than that"), irra comes after object (second item being compared) [35].

Caalaa can come between or after the nouns as shown in the following example:

Manni kee koorra guddaa dha. -----Your house is bigger than mine."

Itto handaaqqoo caalaa kochee nyachuun jaaladha. Or Itto handaaqqoorra kochee caalaa nyachuun jaaladha. Or Itto handaaqqoorra kochee nyachuu caalaan jaaladha. ------I like to eat doro watt more than kitfo.

The descriptors "older" and "younger" are somewhat special cases. Hangafuu is a verb meaning "to be older", while quxusuu is an adjective meaning "younger". They are used as in the examples below:

"My sister is two years older than me." ----- "obboleettiin koo waggaa lama na hangafti." "My sister is two years younger than me." -----"obboleettiin koo waggaa lama quxusuu kooti."

To speak of things being the same, one may use tokkuma ("same"), gosa tokkicha ("the same kind"), or wal fakkaataa ("similar"). Something that is different is adda, and things that are different from each other are adda-adda. Examples are shown below:

These two things are the same-----waantoota lama kunniin tokkuma

These two things are similar-----waantoota lama kunniin wal fakkaataa

These two things are different-----waantoota lama kunniin adda-adda

This one is different-----inni Kun adda

The adverbs ol(i) ("up, above") and gad(i) ("down, below") may be used to compare things as "higher" or "lower", as in:

He is shorter than 1.8 meters. ----- Inni meetira 1.8 (tokko tuqaa saddeet) gadi dha.

He is taller than 1.8 meters. ----- Inni meetira 1.8 oli dha.

3.3 word, phrase and sentence

Word is a single distinct meaningful element of speech or writing, used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed. In linguistics, a word is the smallest element that can be uttered in isolation with objective or practical meaning [27].

Phrase is a small group of words standing together as a conceptual unit, typically forming a component of a clause. Phrase is a group or words that express a concept and is used as a unit within a sentence [28].

Sentence is a set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses. A sentence is a group of words that are put together to mean something. A sentence is the basic unit of language which expresses a complete thought [35].

For the purpose of this study we used the term word for phrases which have max 4 phrase length, phrase for max 16 phrase length and sentence max 30 phrase length.

3.4 Alignment Challenge of Afaan Oromo – English language

Alignment plays a critical role in statistical machine translation by mapping source sentence to target sentence. However, automatic alignment of parallel sentence pair is not a simple task. For most parallel texts, choosing the sentences in one natural language to be the translation of another language is challenging activities. Words may have different combination or alignment that is one-one or one-many or many-one and many-many this makes alignment of words difficult. The figure below shows the alignment properties of English and Afaan Oromo text for both direction:

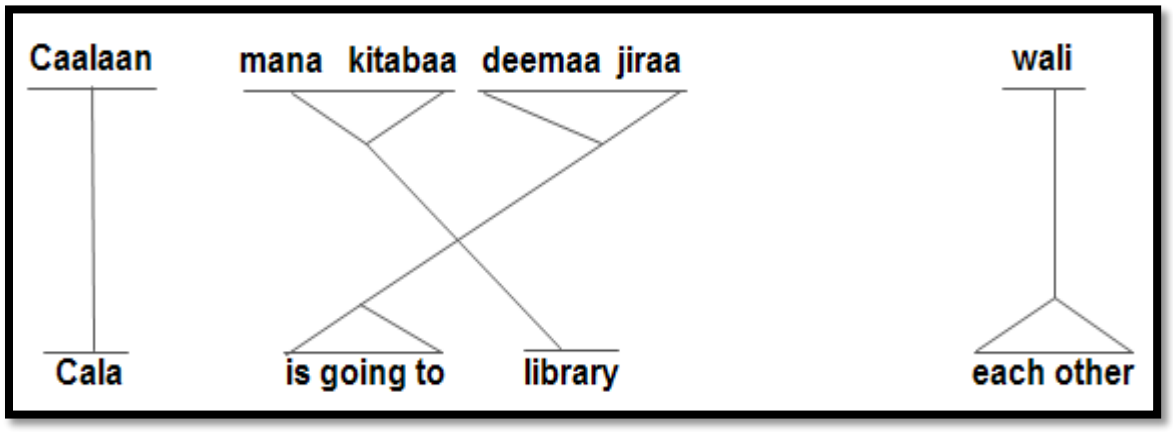

Figure 3.1: Alignments of English and Afaan Oromo sentence

Generally when consider the overview of both languages in some extent they are similar and in another way they are different, especially on number of words they are used in order to construct phrases or sentences, as observed on the examples of the above given sentences that explains the grammars of both languages. This features of the languages makes challenge for the study of machine translation especially for those whose major focuses on alignments.

CHAPTER FOUR

Designing of the MT system

4.1 Corpus preparation

For this study the corpus is collected from different online sources which contain parallel text (English text and Afaan Oromo text). These sources include Megleta Oromia, criminal code, Holly Bible, Constitution of the federal democratic republic of Ethiopia and simple sentences corpus which adapted from [8, 9]. Megeleta Oromia describes about Oromia regional state power and the structure of the power, criminal code describes about the act of crime and the penalties, FDRE constitution describes about the right and responsibilities of people of Ethiopia, and Holly Bible describes the relationship between human and God.

During corpus preparation the following basic activities are performed; we make the length of the sentences balanced from both target and source language by using coma as delimiter to split too long sentences and to combine short sentences together, because, on sentence level alignment, hunalign uses information about length of the sentences for alignment purpose. The following example shows how we made the splitting and merging of the sentences:

Example 1: Splitting too long sentences

If the student will not use the math at a later date, it makes no sense to study Advanced Pure Math.

If the student will not use the math at a later date

It makes no sense to study Advanced Pure Math.

Example 2: Combining short sentences together by separating comma

I have a cat.

It has a nasty temper.

I have a cat, but it has a nasty temper.

Corpus from Megeleta Oromia is difficult because, the source on the web is scanned image format therefore, we edit manually and used for the study.

By performing this it is possible to prepare better size of corpus for sentence level alignment. After preparing the corpus for both languages based on the above method, we used hunalign in order to align the corpus at sentence level. The output of the sentences in some cases they are miss aligned but, it is editable therefore, by editing the output we used for sentence level experiments.

For word level alignment we have used the prepared corpus and MGIZA++ align the corpus using IBM model 1-5. For phrase level alignment we have used the corpus from the source as it is, because, Anymalign align the corpus by using comma, colon, semi-colon and hyphen, therefore we used the corpus from the source as it is in order to keep the punctuation safe.

Generally, based on the alignment properties of tools used for alignment it's possible to prepare large size corpus for SMT.

4.2 Types of the corpus used for the study

For the purpose of the study both simple and complex sentences are used for the experiment. Simple sentence is a sentences containing only one clause with single subject and predicate [36]. Complex sentence a sentence containing a subordinate clause or clauses [36]. The reason why we used both types of sentences is in order to make the corpus balanced in terms of length of sentences to have better language model and translation model.

Types of sentence	Amount
1. Simple sentence	700
2. Complex sentence	5700
Total……..…..6400	

Table 4.1: Summary of corpus size used.

In order to achieve better performance in statistical machine translation using more aligned corpus is recommended. Generally for this study total of 6400 sentences are used (both simple and complex sentences) we used 90% of the corpus for training and 10% for testing randomly based on [1], [7], [8], [9] related study.

4.3 Architecture of the system

This section is about the prototype of the system starting from input corpus until the translation output and the activities performed at each stages.

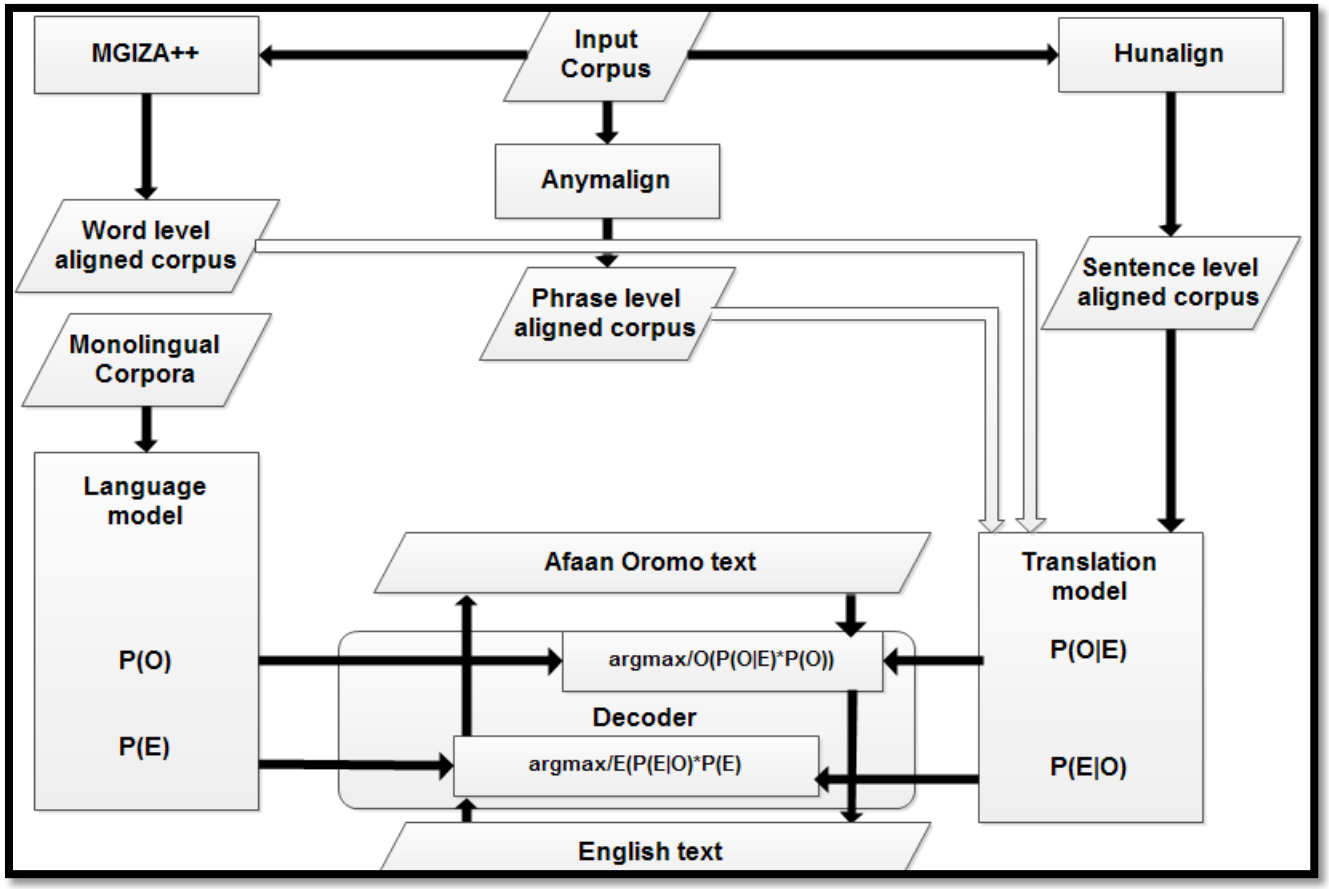

Figure 4.1: Architecture of the system

Corpus was collected from different online source such as FDRE constitution, Megleta Oromia, Criminal code and Holly Bible and prepared by performing different preprocessing tasks such as sentence splitting, sentence merging and true casing. Sentence splitting is the process of splitting two sentences. In order to split the sentences we use coma as delimiters foumureera aniis dadhabeera.r both target and source language. Sentence merging is the process of combining short sentence together. True casing is the process of changing letters form upper case to lower case and the revers, we change the upper case letters to lower case for suitability of the MT system. Then the corpus is ready for the alignment of both target and source languages at word level, phrase level and sentence level using MGIZA++, Anymalign and hunalign tools, respectively. The activities at each level of alignment and language and translation model are discussed as follows:

4.3.1 Word level alignment using MGIZA++

MGIZA++ align the prepared corpus at word level by using IBM models (1-5). The result of the output used for training and testing of the system. The translation and language model then calculate the probability distribution, by using the product of these probability the translation is performed by the decoder from English-Afaan Oromo and From Afaan Oromo-English.

4.3.2 Hunalign

Hunalign, aligns the sentences based on their length and lexical similarity. In order to make the corpus more suitable for the tool we prepared the corpus of both target and source language in to balanced sentences in terms of length. After this the tool aligns the corpus at sentence level by using length of the sentences and lexical similarity. At this stage some of the corpus miss aligned from source language to target language therefore, in order to handle this problem we correct the output manually. Then the output is used for language model and translation model. Then translation model and language model calculate probability distribution for the sentences. By using these probability value the decoder perform the translation from English-Afaan Oromo and From Afaan Oromo-English.

4.3.3 Anymalign

Anymalign is a multilingual sub-sentential aligner. It can extract phrase equivalences from parallel corpora. Its main advantage over other similar tools is that it can align any number of languages simultaneously.

This algorithm align the given corpus at phrase level by using coma and hyphen (, -) respectively as main delimiters or end of line (EOL) to find the phrases of both the source and target language. This two delimiters, comma and hyphen (, -) used in both Afaan Oromo and English languages common delimiter to identify phrases in the sentences, but, another delimiter of phrases in the sentences in both languages are semi colon and colon (; and :). In order to use these marks as additional delimiter we modified the algorithm to find better aligned phrases by including semi colon and colon to algorithm as additional delimiters.

After doing this when we observe the results more of the outputs are aligned at phrase level correctly by following the rule of the algorithm but, some outputs are long phrases this is main

challenge on this stage. Because, it has an effect on the probability of phrase translation table by increasing complexity of the sentences or zero probability.

At the end of the algorithm execution the output is phrase level aligned corpus. By taking this corpus the language model and translation model calculate the probability of phrases. Based on the probability of this values the decoder perform the translation from English-Afaan Oromo and from Afaan Oromo - English.

In order to select the above alignment tools from other tools our criteria is it goes with our objective and it is modified tools which is current tools used for MT research purpose.

4.3.4 Language model

For the language model we used monolingual corpora. 19300corpus used for English and 12200 for Afaan Oromo both are simple and complex sentence. 3-gram model was used based on the nature of the corpus that used for the language model.

4.3.5 Translation Model

For the translation model we used the results of MGIZA++, Anymalign and Hunalign that is word level, phrase level and sentence level aligned corpus respectively.

4.3.6 Decoder

Decoding is a search for the shortest path in an implicit graph [1]. A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability. By following the above procedure the decoder perform the translation process from both directions.

4.3.7 Evaluation

In order to evaluate the performance of the prototype, first we prepare the translated document by the system. Second human translated document which is used as reference translation. By using these two documents BLUE score evaluate the performance of the system.

CHAPTER FIVE

Experiment

After designing, the experiments are conducted based on different length of aligned phrases from both directions (from English-Afaan Oromo and from Afaan Oromo-English). Findings of the experiment and interpretations are presented as follows:

5.1 Experiment I: Experiment done with max phrase length 4 (from English-Afaan Oromo)

In these two experiments we used word level aligned corpus which have maximum phrase length 4 and minimum 1 for the translation process of the input text. The source language is English and the target language is Afaan Oromo (the input text is in English language and the expected output text is in Afaan Oromo language). We use the following 5 paragraph input English text for translation:

Megleta Oromia

Whereas, it has been found proper to stream line operation and organization of urban local governments on the basis of good governance and democratic principle so as to enable them create huge development capacity for the development of the Region and improvement of the living standard of residents; now, therefore in accordance with the provision of Article 49 (3) (a) of the Constitution of the oromia National Regional State, it is hereby proclaimed as follows:

Criminal code

Legal and medical professionals, psychiatrists, different institutions of higher education and professional associations have made significant contributions through the opinions they gave to the enactment of the law. It is mainly on the basis of public opinion that punishments have increased in respect of crimes like rape and aggravated theft. Moreover, the opinions of legal scholars and the laws and experiences of foreign countries have been consulted to enrich the content of the Criminal Code. Finally, one point that must not be left unmentioned is the matter concerning the determination of sentence.

Holly Bible

Went before them, till it came and stood over where the young child was. When they saw the star, they rejoiced with exceeding great joy and when they were come into the house, they saw the young child with Mary his mother, and fell down, and worshipped him: and when they had opened their treasures, they presented unto him gifts; gold, and frankincense, and myrrh. And being warned of God in a dream that they should not return to Herod, they departed into their own country another way.

Simple sentences

Will you wait for me, or do you want to go ahead?

We had finished our homework, and I am tired.

I had originally planned to attend the meeting however now I find I can't.

Derartu likes to drive she finds it relaxing.

Bontu is afraid of the water consequently, she had trouble passing the swimming test.

Sample translation output shown in figure 5.1:

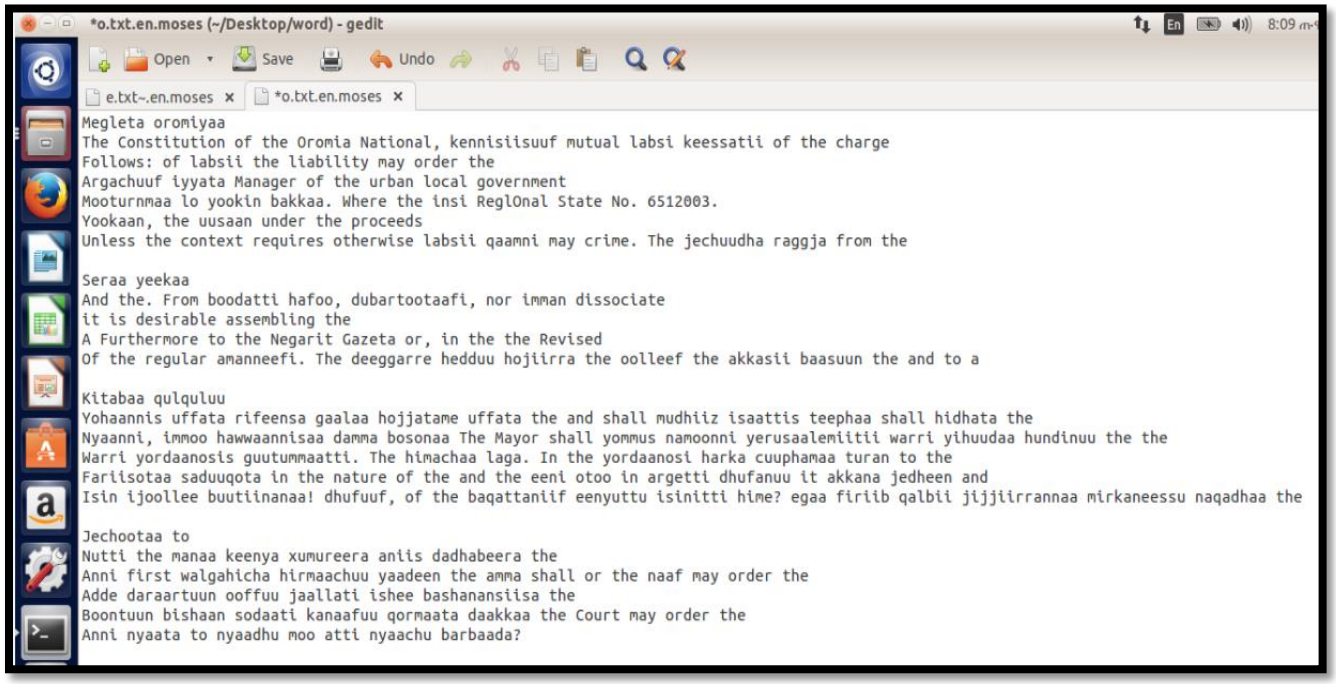

Figure 5.1: Sample translation from English - Afaan Oromo with max phrase length 4

As observed from the above output of translation process some sentences or words are not translated into Afaan Oromo such as "**The constitution of oromia**" in the first paragraph, second line and **"manager of the urban local government"** first paragraph fourth line. This is happen because of the limitation of the sentence alignment during training.

Another basic limitation observed from the output is that, parts of the phrases or sentences are translated. This make readers confused to understand the text this is because, Moses doesn't identifies gender or sex and miss alignment during training of the system due to morphology.

Generally the system translate the given text to the target language (Afaan Oromo) with **21%** BLUE score. The time taken to translate the text is 14 seconds. The following are some reasons for the performance of the system to be 21%.

The first reason is the word correspondence between the languages. Which is the combination of more than one words in Afaan Oromo have single word meaning in English, this increase alignment error and decrease translation performance.

The second reason is position sensitivity of the languages. English language sentence structure Subject Verb Object (**SVO**) but, for Afaan Oromo Subject Object Verb (**SOV**). This is also one of the challenge for the performance to be low. If the source and the target languages have the same structure easy for the system to score better results because, the alignment becomes easy.

The third problem is word level alignment because of the above problems the meaning of the source language text and the target language text not align perfectly. This makes again the performance of the translation low, if the above major problems are handled the quality of the alignment becomes good and so does the translation performance because, sentence alignment and SMT have direct relations.

5.2 Experiment II: Experiment done with max phrase length 4 (from Afaan Oromo-English)

Because of the system works bi-directional, this experiment check the performance of the system with the same corpus used in the experiment I. The same text to translate from source language Afaan Oromo to target language English. We used the following Afaan Oromo text as input for translation.

Megleta oromiyaa

Akkaataan jechichaa hiika biraa kan kennisiisuuf yoo ta'e malee labsi kana keessatii: Magaalaa jechuun jiraattotni haala labsii kana keewwata irratti ibsameen qaama seerummaa argachuuf iyyata dhiheessanii mana Marii Buclhiinsa Mooturnmaa Naannichaatiin yookin qaama bakkaa bu'a insi kennameefiin kan murtaa'eef yookaan labsiin kun ragga'uusaan dura magaalaa mana qopheessaa qabu ta'ee akkaataa labsii kanaan qaamni seerummaa kan raggja'eef jechuudha.

Seraa yeekaa

Dhimmi callisaan bira darbamuu hin qabne kan biroo, gochoonni barmaatilee boodatti hafoo ta’an dubartootaafi daa’imman irratti miidhaafi dararama olaanaa kan geessisan ta’uu isaanii Seerrii Adaba Yakkamaa duraanii beekamtii kennuu dhabuu isaati. Dhugaadha; aadaan ummattootaa akka kabajamu heerri mootummaa ni tuma. Barmaatilee miidhaan isaanii saayinsiin mirkanaa’e garuu Heerri Mootummaa hin deeggaru. Akkasumas, seerri ummattoonni itti hin amanneefi hin deeggarre yeroo hedduu hojiirra waan hin oolleef seera akkasii baasuun faayidaa akka hin qabne dhugaadha.

kitabaa qulquluu

Yohaannis uffata rifeensa gaalaa irraa hojjatame uffata ture; mudhiiz isaattis teephaa hidhata ture; nyaanni isaa immoo hawwaannisaa fi damma bosonaa ture. Yommus namoonni Yerusaalemiitii fi warri Yihuudaa hundinuu, warri naannoo Yordaanosis guutummaatti gara isaa himachaa, laga Yordaanosi keessatti harka isaatti cuuphamaa turan. Yommuu Fariisotaa fi Saduuqota keessaa baay'eeni isaanii otoo gara isaa dhufanuu argetti garuu akkana jedheen; Isin ijoollee buutiinanaa! dheekkamsa dhufuuf jiru jalaa akka baqattaniif eenyuttu isinitti hime? Egaa firiib qalbii jijjiirrannaa mirkaneessu naqadhaa.

Jechootaa addaa

Nutti Hojii manaa keenya xumureera aniis dadhabeera.

Anni jalqaba walgahicha hirmaachuu yaadeen ture ta'us amma akka hin dandeenye naaf gale.

Adde daraartuun ooffuu jaallati ishee bashanansiisa.

Boontuun bishaan sodaati kanaafuu qormaata daakkaa darbuu hin dandeenye.

Anni Nyaata kana nyaadhu moo Atti nyaachu barbaada?

The sample output is shown in the figure 5.2:

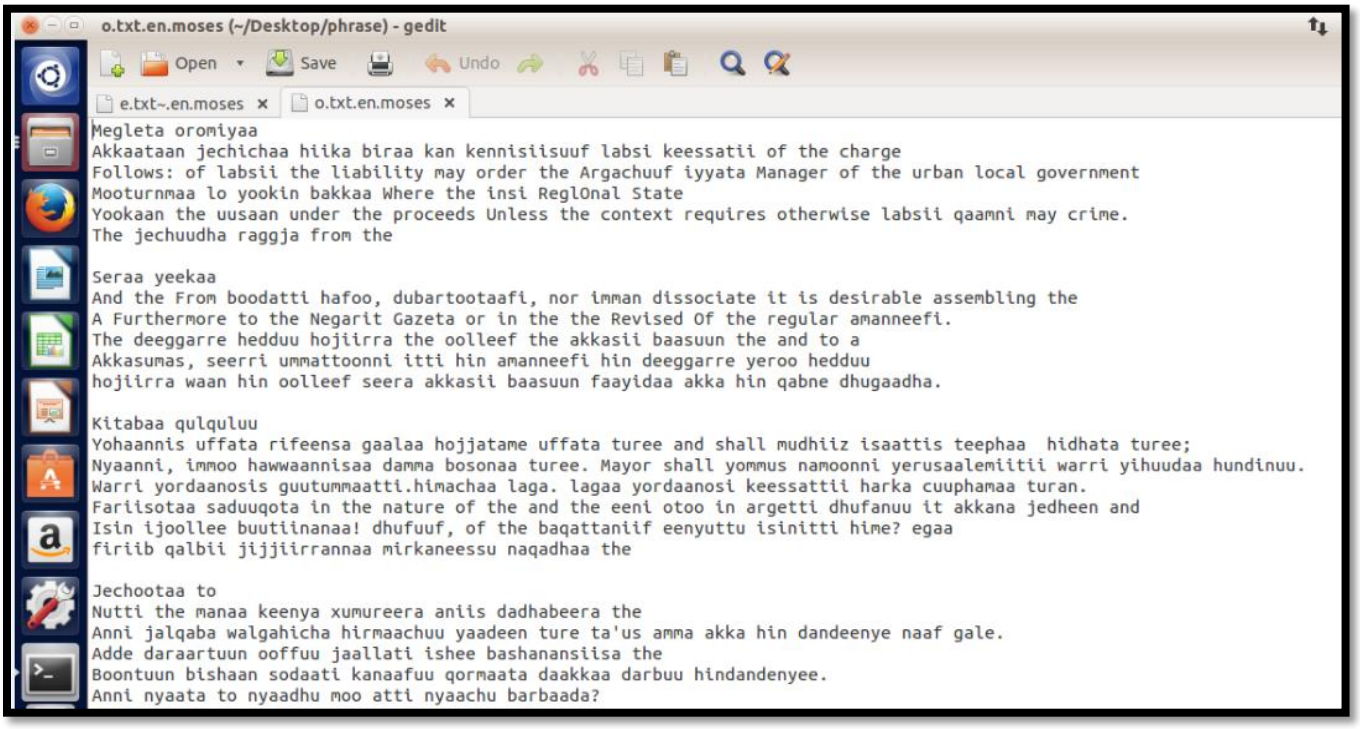

Figure 5.2: Sample translation from Afaan Oromo - English with max phrase length 4

As shown from the above output some words are jumped without translation example (Akaataa first paragraph second line, Akkasumas second paragraph fourth line and etc.). This happens because of alignment problem. Example for Afaan Oromo text "**mana kitabaa**" similar meaning with "**library**" in English, in this case the translation is miss aligned because, the first text constructed from two word but, the second word which is the same meaning for the first text is constructed from single word (word correspondence). **42%** of BLUE score recorded from the above translation process.

From the above two experiments (I and II) we consider that the system achieve better performance when Afaan Oromo is source language, because of, alignment probability of the words. When the system train by taking Afaan Oromo as source language it gates more number of words aligned

than the previous experiment. This makes the performance of the translation better when compare with experiment I.

5.3 Experiment III: Experiment done with max phrase length 16 (from English-Afaan Oromo)

The following two experiments shows the results of the translated text after the system is trained by phrase level aligned corpus with max phrase length 16 and min 4. The phrase used for these two experiments are longer than word level aligned corpus and shorter than sentence level aligned corpus. The first experiment conducted by taking English as source language and Afaan Oromo as target language. We used the same English input text as Experiment I, the result of the experiment shown in the figure 5.3:

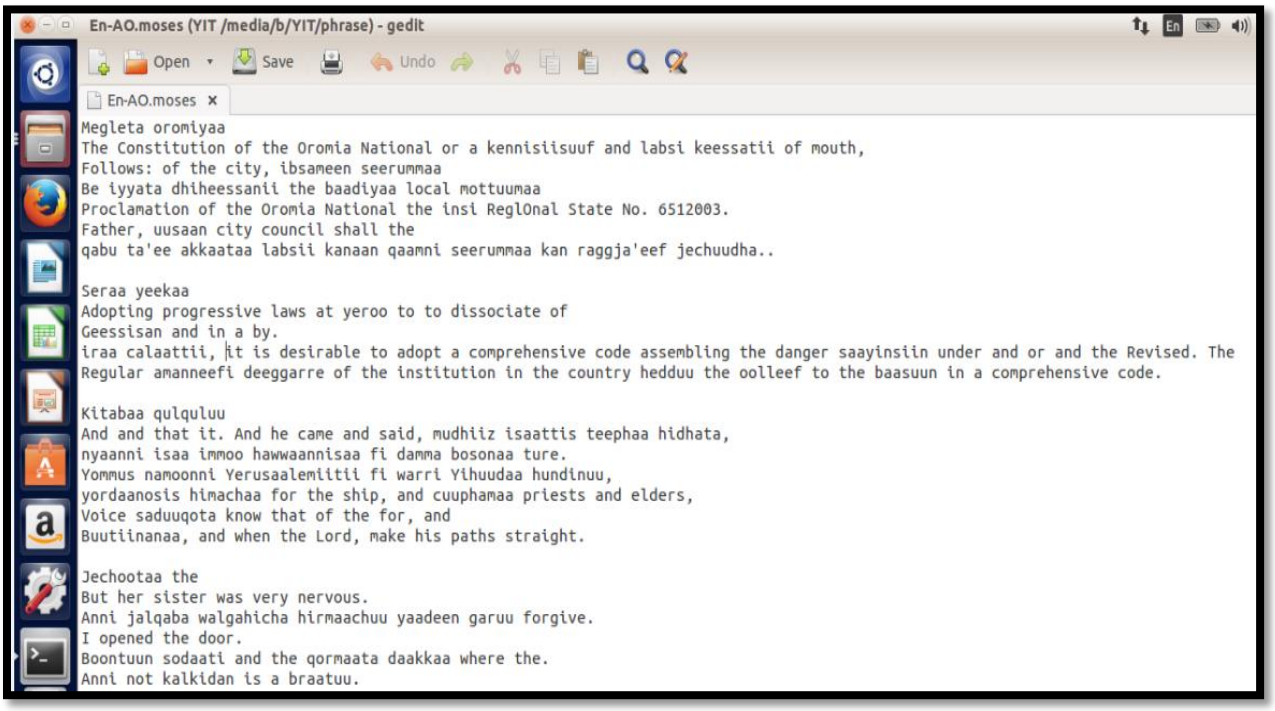

Figure 5.3: Sample translation from English – Afaan Oromo with max phrase length 16

The performance of the above translation is **27%** BLUE score. When we compare this result with experiment I the performance of the translation system is improved by **6%.** The reason why the result improved is, because of phrase level alignment.

From the above experiments (I and II) one major reason for the performance of the system to be low is the problem of language correspondence (sentence or phrase formation between both

languages), therefore, this problem is handled by the phrase level alignment. Even if the word correspondence handled, the structure of the languages and the length of the phrases are a major factors for the performance of the translation system to be only 27%. As phrase length increase the probability value of the phrases decrease this increase non-aligned phrases. Example:

Addisuun hojjetaa dha | Addisu is an employee (21 character)…….0.7

Akkaataan jechichaa hiika biraa kan kennisiisuuf yoo ta'e malee labsi kana keessatii | Unless the context requires otherwise in this Proclamation City means a community of (84 character)……0.2

As the length of the character or number of word used in phrase increase the probability value decrease this increase non-aligned phrase.

5.4 Experiment IV: Experiment done with max phrase length 16 (from Afaan Oromo - English)

To conduct this experiments first the prepared corpus is aligned at phrase level with multilingual aligner (Anymalign). After the alignment is done to translate the prepared text the system trained by this corpus (phrase level aligned corpus) and finally by giving the input (Afaan Oromo text same with experiment II) the translation process is done. Sample output of the experiment is presented in figure 5.4:

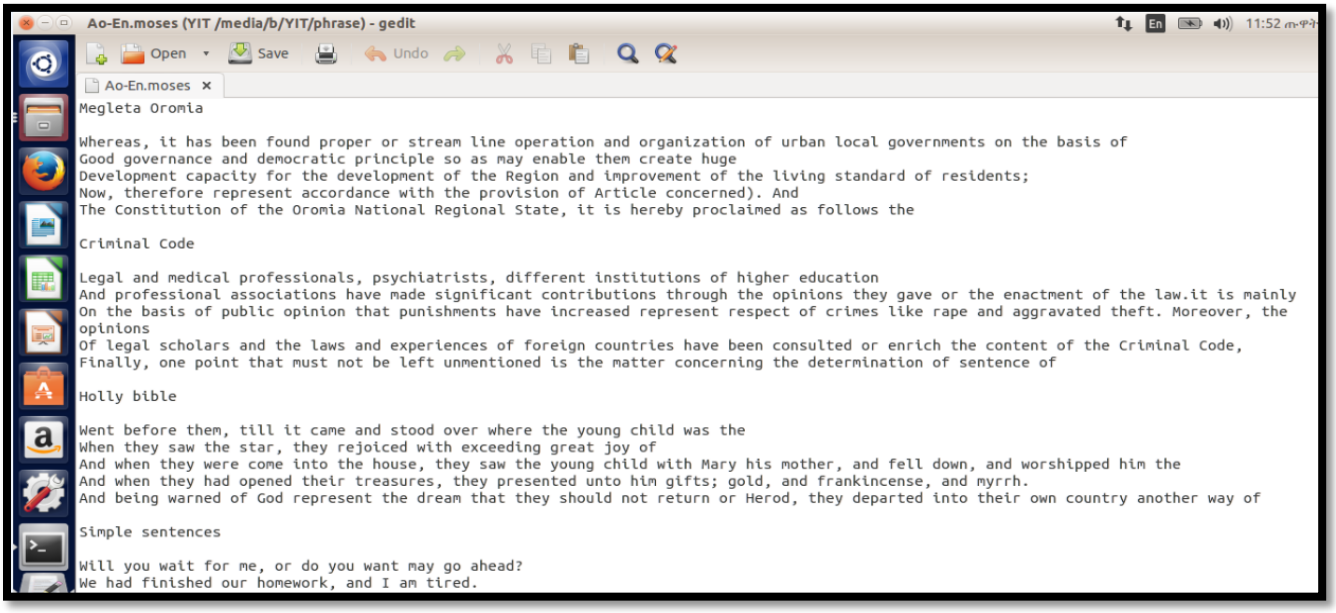

Figure 5.4: Sample translation from Afaan Oromo-English with max phrase length 16

From the output of the experiment above some of the sentences translated when we compare with experiment II. The BLUE score recorded for this experiment is **47%**. When we compare this result with experiment II result it is better, the reason is the same with experiment III that is the phrase level alignment handle the word correspondence. The structure of both the source and the target language is different, this makes the translation performance low. If the source and the target languages are the same, the result of the system is better than this.

5.5 Experiment V: Experiment done with max phrase length 30 (from English - Afaan Oromo)

The following two experiments show the results of the translated text after the system is trained by sentence level aligned corpus with max 30 and min 20 phrase length done by hunalign. The phrase length used for these Experiments are longer than phrases on word level aligned and phrase level aligned corpus. The first experiment conducted by taking English as source language and Afaan Oromo as target language and we used same input Afaan Oromo text as experiment I the result of the experiment shown in the figure 5.5:

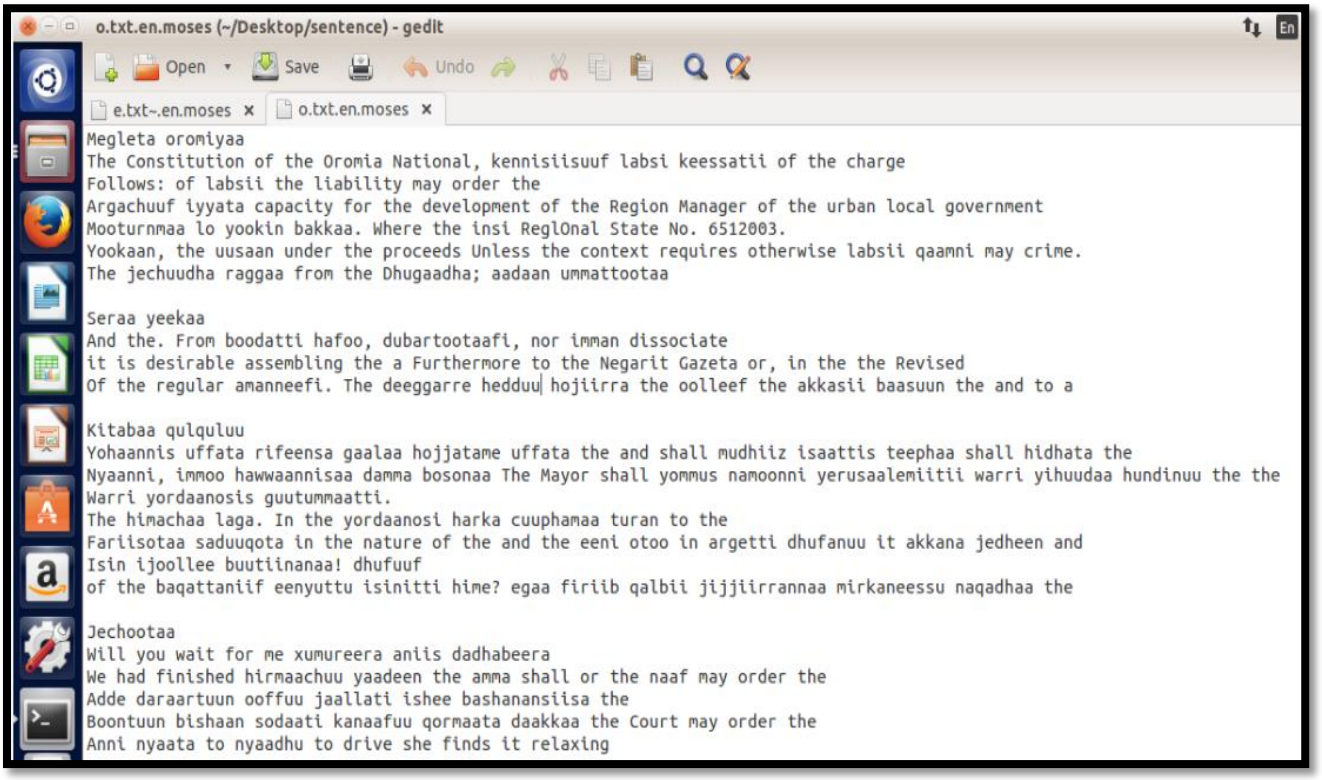

Figure 5.5: Sample translation from English-Afaan Oromo with max phrase length 30

When we compare the result of this experiment again with the results of experiment above some of the sentences in the paragraph are not translated. The reason is in order to handle the alignments problem of the corpus we used sentence level aligned corpus for translation model. This makes alignments difficult, because, complexity of the sentence becomes high this resulted to poor translation performance. The structure of both the target and source language is also a factor for the performance to be low.

For this experiment **18%** BLUE score recorded. When we consider all the experiments better BLUE score is achieved or recorded when English is used as target language and Afaan Oromo as source language. This is because of alignment quality is better when English is used as target language whether we used word level alignment, phrase level alignment and sentence level alignment during the training of system.

5.6 Experiment VI: Experiment done with max phrase length 30 (from Afaan Oromo-English)

This experiment is the same with experiment V, the difference is, in this case the source language is Afaan Oromo and the target language is English. The translation model are trained by using sentence level aligned corpus like experiment V. We use the same English text as input for translation like Experiment II the result of the experiment is as shown in the figure 5.6:

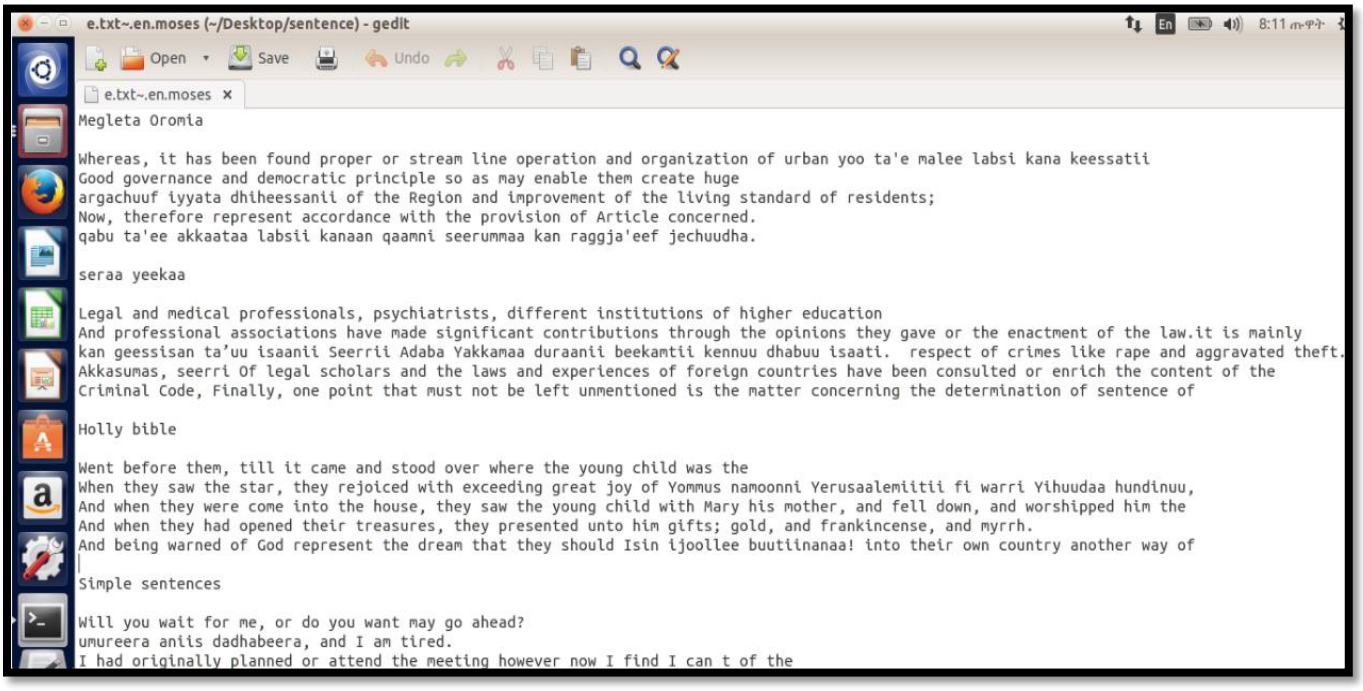

Figure 5.6: Sample translation from Afaan Oromo-English with max phrase length 30

When we consider the result of this experiment with above experiment some of the texts are jumped without translation. The BLUE score recorded for this experiment is **35%.** This indicate that when the corpus aligned at sentence level and the translation model is trained with this corpus the complexity of the sentences becomes high. This makes the ratio of zero probability increase (non-aligned corpus increase) therefore, this condition affects the translation performance. The structure of the source and target language also another factor for the performance to be low.

Based on the result of the experiment, in order to handle the word correspondence of the sentences which is basic challenge for alignments of both target and source languages, phrase level alignment with max phrase length 16 is the optimal one, than word level alignment or sentence level alignment, in order to handle the alignment problem of both target and source language.

5.7 Result and discussion

The main purpose of this study is to conduct experiment on bi-directional English-Afaan Oromo, statistical machine translation to explore an optimal alignment level for better performance of statistical machine translation. Different experiments are conducted from English-Afaan Oromo and from Afaan Oromo-English language. The result of the experiments shown in the table 5.1:

Max and Min length of phrases	Result of experiment in BLUE from both directions	
	English – Afaan Oromo	Afaan Oromo – English
Max 4 and min 1	21%	42%
Max 16 and min 4	**27%**	**47%**
Max 30 and min 20	18%	35%

Table 5.1: Summary of Experiment result.

As shown from the above result summary, an optimal alignment is phrase level alignment when the max phrase length is 16 and min is 4 which record 27% and 47% BLUE score from English-Afaan Oromo and From Afaan Oromo-English respectively.

In order to achieve better result the corpus is aligned at phrase level by using Anymalign algorithm. This decreases the number non-aligned phrases in the corpus and increase the number of aligned phrases at phrase translation table. This makes the translation performance better. Some outputs

of aligned phrases are too long this affect the performance of phrase translation table which is a backbone for statistical machine translation therefore, this is one challenge for this study. The other challenge is, there is no hybrid alignment that handle alignment varieties.

For the starting of this study as we discussed in the statement of the problem two researches [8, 9], which focus on machine translation on Afaan Oromo language. From these two works mainly the first work which focus on SMT for Afaan Oromo is related with this study specially the approach the author followed for the study that is statistical approach. In this study [8], the author used word level alignment for the corpus and he made experiments and finally he achieves 17% of BLUE score translation performance.

The major activity in this study is, we used not only word level alignment but also phrase level and sentence level alignment, because, the structure of both the target and source languages word correspondence is not only one-one rather it includes one-many, may-one, many-many. In order to handle this problem we use phrase level alignment by using Anymultilingual aligner algorithm. Identifying an optimal alignment level for better performance of SMT is basic strength of this study. This level of alignment is only tested in Afaan Oromo and English language pair but, not tested with other language pair.

Generally the translation performance of this study on average **32%** BLUE score. When we compare this result with the previous research work [8] **17%**. The translation performance of this study is better because, the activities related with alignment of prepared corpus (word level, phrase level and sentence level) which is basic challenge for SMT, is studied in this study to overcome alignment challenges.

CHAPTER SIX

Conclusion and recommendation

6.1 Conclusion

Alignment of the corpus and statistical machine translation have strong relation because, in order to translate text SMT learns from properly aligned corpus. In this study we explored an optimal alignment by considering the source and target language of the study (Afaan Oromo and English language). In order to explore the alignment first we studied the sentence structure of both Afaan Oromo and English language. Then we identify the word correspondence between languages is one-one, one-many, many-one and many-many. Then by aligning the corpus at different level of alignment (word level, phrase level and sentence level) we conduct experiments. We identify phrase level alignment is an optimal level of alignment for better SMT performance for Afaan Oromo and English language pair.

The design process of bi-directional English-Afaan Oromo statistical machine translation involves collecting English-Afaan Oromo parallel corpus. The corpus collected from freely available on-line sources such as Ethiopian constitution, criminal code, Megleta Oromoia, holly bible and simple sentences adapted from [8,9]. Corpus preparation involves activities of preprocessing the corpus such as sentence splitting and true casing. Aligning the prepared corpus by considering the structure of both languages. MGIZA++ used for word level alignment, multilingual aligner (Anymalign) used for phrase level alignment and Hunalign used for sentence level alignment. Moses for mere mortal used for translation process which integrate all necessary tools for machine translation such as IRSTLM, MGIZA++ and decoder.

After designing in order to identify the optimal alignment different experiments are conducted under taking level of alignment as major category for the experiments. Based on this, the study identify the phrase level alignment is an optimal level of alignment for the study by scoring 27% and 47% BLUE score from English-Afaan Oromo and from Afaan Oromo-English respectively. The reason for this alignment to be optimal is, it contribute more phrases for phrase translation table than the rest level of alignments for better performance of statistical machine translation.

Identifying the optimal level of alignment by conducting different experiments which used to enhance the statistical machine translation performance is the strength for this study. From the findings of this study phrase level alignment is an optimal one from the rest level of alignment for English and Afaan Oromo language pair, but this is not tasted with other pair of language. During phrase level alignment, some output of the algorithm is too long this affect phrase translation table. This is one challenge that we are faced in this study. Another challenge is there is no hybrid alignment that handle varieties of alignment.

Generally this study conclude that phrase level aligned corpus improve the performance of statistical machine translation, when the source and the target languages are English and Afaan Oromo.

6.2 Recommendation

Statistical machine translation is one of corpus based approach for translation. It trains and translate based on the corpus prepared for training. Generally we would like to recommend the following points for further works:

- ✓ On phrase level alignment we use coma (,), hyphen (-), semi colon (;) and colon (:) in order to find and align phrases but, even if we used those character the output of some phrases length is long this can affect the translation performance therefore, if this condition is handled better result can be achieved.
- ✓ In both the source and target languages of this study there is varieties of alignment exist such as one-one, one-many, many-one and many-many. If hybrid approach that handle this alignment is developed better SMT result recorded.
- ✓ Better results can be achieved by using the corpus with proper alignment used for training the system. So, by increasing the size of the training data set that properly aligned at phrase level one can develop a better bi-directional English-Afaan Oromo machine translation.
- ✓ Most of the corpus used for this study is collected from legal document, if the corpus prepared from different discipline better result can be recorded.

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. D. Okpor, "Machine Translation Approaches: Issues and challenges," IJCSI International Journal of Computer Science, Vol. 11, No 2.Issue 5, pp. 159-165, 2014.

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

[5] 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.

[6] 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.

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

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

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

[10] N. Brad , "LLC dba Diplomatic Language Services," 2017. [Online]. Available: http://dlsdc.com/blog/machine-translation-advantages-and-disadvantages/.[Accessed 15/5/2017 may 2017].

[11] C. Kothari, Research methodology, india: New age international (p) limited, publishers, 2004.

[12] R. M. Steven and M. R. Gary , "Experimental Research Method," in Experimental Research Methods, memphis, wayane, 2003, p. 25.

[13] Mulu Gebreegziabher Teshome and Laurent Besacier, "Preliminary experiments on English-Amharic statistical machine translation".

[14] H. J. W, "Machine translation: a brief history," in Concise history of the language sciences: from the Sumerians to the cognitivists., K. F. E and A. E. R, Eds., Oxford, Pergamon press, 1995, pp. 445-460.

[15] H. J. W, "Machine translation: a brief history," Concise history of the language sciences: from the Sumerians to the cognitivists, pp. 431-445, 1995.

[16] A. Douglas , B. Lorna , M. Siety , H. R. Lee and S. Louisa , Machine Translation An Introductory Guide, London: NCC Blackwell Ltd, pp.234, 1994.

[17] S. Michel and P. Pierre , "Bilingual sentence alignment balancing robustness and accuracy," Centre for Information Technology Innovation (CITI), pp. 135-144.

[18] B. Fabienne and F. Alexander , "Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora," Institute for Natural Language Processing, pp. 81-89, August 2010.

[19] S. R. Jason, Q. Chris and T. Kristina , "Extracting parallel sentences from comparable corpora using document level alignment," HLT '10 Human Language Technologies, pp. 403-411, 02 June 2010.

[20] M. C. Robert, "Fast and Accurate Sentence Alignment," Inistitute of natural language processing focus machine translation pp. 135-144, 2002.

[21] S. Andr´e , "A survey on parallel corpora alignment," MI-STAR, vol. 12, pp. 117-128, 2011.

[22] T. Liang , W. Fai and C. Sam , "Word Alignment Using GIZA++ on Windows," University of Macau, Macau, pp. 369-372, 2010.

[23] A.-R. Sadaf , F. Mark , L. Patrik , N. Sandra and S. Rico , "Extrinsic Evaluation of Sentence Alignment Systems," in LREC Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS), Istanbul, Turkey, 2012.

[24] Adrien Lardilleux and Yves Lepage, International Conference on Recent Advances in Natural Language Processing, september 2009. [Online]. Available: https://anymalign.limsi.fr/. [Accessed 20 february 2017].

[25] A. Ibrahim and S. S. Ibrahim , "Intelligent hybrid man-machine translation," Alexandria, 2014.

[26] Z. Yue , "Chinese-English Statistical Machine," MSC thesis British, Oxford University, 2006.

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

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

[29] 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.

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

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

[32] W. T. Abire, "Passivization in Afaan Oromoo," Academy of Ethiopian Languages and Cultures, pp. 10-18, june 2012.

[33] A. Raga and S. Adola, "Homonymy as a barrier to mutual intelligibility among speakers of various dialects of Afan Oromo," Journal of Language and Culture, vol. 3(2), no. 2141-6540, pp. 32-43, 2012.

[34] T. Debela, "A rule base afan Oromo grammar checker," proceedings of the International Journal of Advanced Computer Science and Applications, 2011.

[35] A. B. Dhinsaa, Sanyii : jechaafi caasaa isaa / Afan Oromo Word and Its Structure, Finfinnee: Addunyaa Barkeessaa, 2013.

[36] L. Dana , Book " The Hungrer Games" Sentences : Simple , Compound and complex compound sentences, 2014.

[37] S. Anoop , "Generative Model of Word Alignment," in Natural Language Processing, Simon Fraser , Simon Fraser University, 2016, pp. 345 - 360.

Appendices

Appendix I: URL for sources of the corpus

1. https://www.unodc.org/cld/document/eth/2005/the_criminal_code_of_the_federal_democratic_republic_of_ethiopia_2004.html Criminal code English version
2. http://www.abyssinialaw.com/codes-commentaries-and-explanatory-notes?download=1208:fdre-criminal-code-afaan-oromo-version
3. https://www.google.com/search?client=opera&q=heera+motummaa+naannoo+oromiyaa&sourceid=opera&ie=UTF-8&oe=UTF-8#q=heera+mootummaa+feederaalawaa+itoophiyaa+pdf Heera mootumaa Ethiopia
4. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=17&cad=rja&uact=8&ved=0ahUKEwjo Ethiopian constitution
5. https://www.lds.org/scriptures/nt/matt/1?lang=eng Holly bible English version
6. https://app.box.com/s/08vsopn8cwb63rv32yth Holly bible Afaan Oromo version.
7. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwj9h5nt5fvTAhVM82MKHQ_MAdUQFggxMAI&url=http%3A%2F%2Fextwprlegs1.fao.org%2Fdocs%2Fpdf%2Feth153469.pdf&usg=AFQjCNGbexcFJpSJBb9jyfcIl7mXO4zhUA Megleta Oromia both in English and Afaan Oromo.

Appendix II: sample of word level aligned corpus

Afaan Oromo	English
Namni kamiyyuu	Whoever incites another
Sababa babal'ina gocha	Where the case is more serious
Meeshaa yookiin kaappitaala	knowingly supplies
Tooftaa akaakuu biro	in any other way
Abbaan taayitaa yookiin hojjetaan	Any official or employee of an authority who
Namni kamiyyuu itti yaadee	Whoever intentionally brings
Yakkichi kan raawwatame humnaan	Where the crime is committed
Mi'oota, oomishoota yookiin	the importation exportation storage or
Bu'aa qabeenyaa	the exploitation
Bineeldota	the settlement
Bojii hayyama mootummaa	a monopoly whether granted
Caasaa baankii	the organization of State banks
Gochoota keewwata kanaan	Where one of the acts in this Article
Haalawwan Yakkicha Cimsan	Aggravation to the Crime
Yakki keewwata xiqqaa tokko	Where the crime specified in sub-article 1
Hojjechuu	Making
Sobatti Jijjiiruu	Forgery.
Gatii Gadi Buusuu	Debasing.
Tilmaama yaada Naannessuu	Presumption of Intent to Utter.

Appendix III: sample of phrase level aligned corpus

Afaan Oromo	English
Yakki tokko namoota lakkoofsi isaanii tokkoo ol ta’aniin gamtaan yommuu raawwatame,	Where two or more persons commit a crime in concert,
Barattootni taphachaa jiru.	The students are playing.
Getnet kubbaa qaba.	Getnet has a ball.
Inni haadha isaa jaallata.	He loves his mother.
keewwata xiqqaa (3)jalatti kan ibsame bu'ura taasisudhan qamnii sababa quubsaadhan seerummaa akka hin kennamne yoo murtaa' e	Sub Article (3) of this Article declines the request for good reason. it shall mention the reasons
Simboon haadha manaa isaati.	Simbo is his wife.
Addisuun hojjetaa dha.	Addisu is an employee.
Yohaannis shaayee dhugaa jira.	John is drinking tea.
Addisuun waggaa lama dura fuudhe.	Addisu has been married for two years.
Akkaataan jechichaa hiika biraa kan kennisiisuuf yoo ta'e malee labsi kana keessatii	Unless the context requires otherwise in this Proclamation City" means a community of
Almaaz kubbaa saaphanaa taphachuu jaallatti.	Almaz likes playing volleyball.
Gaaddiseen kubbaa milaa taphatti	Gadise plays football.
keewwata xiqqaa (3)jalatti kan ibsame bu'ura taasisudhan qamnii sababa quubsaadhan seerummaa akka hin kennamne yoo murtaa' e	Sub Article (3) of this Article declines the request for good reason. it shall mention the reasons
Seerri haqame yeroo hojiirra turetti yakkoota raawwataman irratti seerri kun erga ragga’ee booda murtiilee kennaman irratti haala murtii tarkaanfilee	Upon the coming into force of this Code measures
Namni kamiyyuu Itoophiyaan alatti lammii Itoophiyaarratti yookiin	This Code shall also apply to any person who has committed a crime outside Ethiopia against an Ethiopian national or

Appendix IV: sample of Sentences level aligned corpus

Afaan Oromo	English
Galmi Seera Yakkaa yakki akka hin raawwatamne ittisuu yommuu ta'u kanas kandhugoomsu waa'ee gochoonni yakkaafi adabbii isaanii dursee akeekkachiisa kennuun, yommuu akeekkachiisichi gahaa hin taanettis raawwattoonni yakkaa adabamanii yakka biroo raawwachuurraa akka of qusataniifi kanneen birootiif barumsa akka ta'an yookiin akka sirreeffaman taasisuun yookiin yakkoota dabalataan akka hin raawwanneef tarkaanfilee akka isaanirratti fudhataman taasisuunidha.	Republic of Ethiopia is to ensure order, peace and the security of the State, its peoples, and inhabitants for the public good. It aims at the prevention of crimes by giving due notice of the crimes and penalties prescribed by law and should this be ineffective by providing for the punishment of criminals in order to deter them from committing another crime and make them a Lesson to others, or by providing for their reform and measures to prevent the commission of further crimes.
Seerri kun akka ragga'u erga taasifamee booda raawwatichi seerichi ragga'uu isaatiin dura yakka raawwateef yommuu itti murtaa'u, yeroo yakkicha raawwatetti seera hojiirra ture caalaa seerri kun adabbii kan isaaf salphisu yoo ta'e adabbiin seera kana keessatti tumame isarratti ni raawwatama. Manni murtichaa seerri kun irra caalaa kan wayyu ta'uu isaa kan murteessu tokkoon tokkoon dhimmaa irratti tumaalee seeraa rogummaa qaban madaaluun ta'a.	Where the criminal is tried for an earlier crime after the coming into force of this Code, its provisions shall apply if they are more Favorable to him than those in force at the time of the commission of the crime. The Court shall decide in each case whether, having regard to all the relevant provisions, the new law is in fact more favorable.
Hiika Akkaataan jechichaa hiika biraa kan kennisiisuuf yoo ta'e malee labs;i kana keessatii: Magaalaa" jechuun jiraattotni haala labsii kana keewwata irratti ibsameen qaama seerummaa argachuuf iyyata dhiheessanii mana Marii Buclhiinsa Mooturnmaa Naannichaatiin yookin qaama bakkaa bu'a insi kennameefiin kan murtaa'eef yookaan labsiin kun ragga'uusaan dura magaalaa mana qopheessaa qabu ta'ee akkaataa labsii kanaan qaamni seerummaa kan raggja'eef jechuudha.	Definitions Unless the context requires otherwise in this Proclamation City" means a community of residents incorporated as a city by the Regional Executive Council or a delegated body in accordance with article 4 of this Proc1amation Urban Local Government means the administration of self-rule by the cities in the Region after acquiring legal personality
Hiddi dhaloota Yesuus Kiristoosi isa sanyii Daawiti, sanyii Abraahami ta'e sanaa kana: Abrahaam Yisaaqin dhalfatea; Yisaaq Yaaqoobin dhalfate;Yaaqoob Yihuudaafii obboleey-yan isaa dhalfate;	The book of the generation of Jesus Christ, the son of David, the son of Abraham. Abraham begat Isaac; and Isaac begat Jacob; and Jacob begat Judas and his brethren;