Table of Contents....i
List of Tables....iii
List of Figures....iv
Acronyms and Abbreviations....v
CHAPTER ONE....1
1. INTRODUCTION....1
1.1. Introduction....1
1.2. Background....1
1.3. Statements of the Problem....2
1.4. Objectives of the Study....3
1.4.1. General Objective....3
1.4.2. Specific Objectives....3
1.5. Methodologies....3
1.5.1 Literature Review....3
1.5.2. Data Collection....3
1.5.3. Tools and Techniques....4
1.5.4. Evaluation....4
1.6. Scope and Limitations of the Study....4
1.6.1. Scope of the Study....4
1.6.2. Limitations of the Study....4
1.7. Contribution of the Study....5
1.8. Organization of the Thesis....5
CHAPTER TWO....6
2. WOLAYTTA LANGUAGE....6
2.1. Introduction....6
2.2. Overview of Wolaytta Language....6
2.3. Morphology....6
2.3.1. Morphological Analysis....8
2.3.2. Morphological Synthesis....8
2.4. Morphology of Wolaytta....9
2.4.1. Personal Pronouns....10
2.4.2. Subject Verb Agreement....10
2.4.3. Nouns....11
2.4.4. Gender....11
2.4.5. Number....11
2.5. Wolaytta Language Writing System....13
2.6. Wolaytta Language Sentence Structure....14
2.7. Articles....14
2.8. Punctuation Marks....14
2.9. Conjunctions....15
CHAPTER THREE....16
3. LITERATURE REVIEW....16
3.1. Introduction....16
3.2. Machine Translation (MT)....16
3.3. Approaches of Machine Translation....17
3.3.1. Statistical Machine Translation (SMT)....17
3.3.2. Rule Based Machine Translation (RBMT)....20
3.3.3. Example Based Machine Translation (EBMT)....24

3.3.4. Hybrid Machine Translation (HMT).....25
3.3.5. Neural Machine Translation (NMT).....25
3.4. Evaluation of Machine Translation.....26
3.5. Related Works.....28
3.5.1. English–Afaan Oromo Machine Translation: An Experiment Using Statistical Approach..28
3.5.2. Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus.....29
3.5.3. Preliminary Experiments on English-Amharic Statistical Machine Translation (EASMT).30
3.5.4. Bidirectional English–Afaan Oromo Machine Translation Using Hybrid Approach.....31
3.5.5. English-Tigrigna Factored Statistical Machine Translation.....32
3.5.6. Bidirectional Tigrigna-English Statistical Machine Translation.....33
CHAPTER FOUR.....34
4. DEVELOPMENT OF ENGLISH-WOLAYTTA SMT.....34
4.1. Introduction.....34
4.2. Architecture of the English-Wolaytta SMT.....34
4.3. Corpus Collection and Preparation.....36
4.3.1. Preliminary Preparation.....36
4.3.2. Bilingual Corpus.....37
4.3.3. Monolingual Corpus.....37
4.3.4. Language Model.....37
4.3.5. Translation Model.....39
4.3.6. Decoding.....40
4.4. Software's.....40
CHAPTER FIVE.....42
5. EXPERIMENT.....42
5.1. Introduction.....42
5.2. Experiment.....42
5.2.1. Experiment-I: Unsegmented Corpus Set.....42
5.2.2. Experiment-II: Segmented Corpus Set.....43
5.3. Discussion.....44
CHAPTER SIX.....46
6. CONCLUSION AND RECOMMENDATION.....46
6.1. Conclusion.....46
6.2. Recommendation.....47
7. References.....48

List of Tables

Table 2.2. Wolaytta Morphological Structure
Table 2.3. Wolaytta masculine and feminine
Table 2.4. Wolaytta nouns singular and plural
Table 2.5. 2nd class Wolaytta nouns singular and plural
Table 2.6. 3rd class Wolaytta nouns singular and plural
Table 2.7. 4th class Wolaytta nouns singular and plural
Table 2.8. Wolaytta language alphabet letter
Table 2.9. Wolaytta language conjunctions
Table 4.1. Bilingual sentences
Table 4.2. Monolingual sentences
Table 5.1. Summary of the total sentences
Table 5.2. Unsegmented experiment sets
Table 5.3. Segmented experiment sets

List of Figures

Figure 2.1. Wolaytta language sentence structure
Figure 3.1. Different levels of analysis in an MT system
Figure 3.2. Direct MT system
Figure 3.3. Transfer MT system
Figure 3.4. Vauquois MT system
Figure 4.1. Architecture of SMT
Figure 5.1. Unsegmented experiment BLEU scores
Figure 5.2. Segmented experiment BLEU scores

Acronyms and Abbreviations

BLEU: Bilingual Evaluation Understudy

EBMT: Example Based Machine Translation

EM: Expected Maximization

FDRE: Federal Democratic Republic of Ethiopia

HMT: Hybrid Machine Translation

LM: Language Model

METEOR: Metric for Evaluation of Translation with Explicit Ordering

MT: Machine Translation

SMT:Statistical Machine Translation

NMT:Neural Machine Translation

NIST: National Institute of Standards and Technology

NISTLM: National Institute of Standards and Technology Language Model

NLP: Natural Language Processing

RBMT: Rule Based Machine Translation

SMT: Statistical Machine Translation

SOV: Subject-Object-Verb

SVO: Subject-Verb-Object

TM: Translation Model

WER: Word Error Rate

CHAPTER ONE

1. INTRODUCTION

1.1. Introduction

This chapter gives general information about the thesis as a whole. It gives the general background of the study, the statement of the problem, the general and specific objectives of the study, the methodologies which are literature review, data collection, development tools & evaluation and scope, limitation and contribution of the study.

1.2. Background

Language is one of the fundamental aspects of human behavior and a basic component in daily activities[1]. It is the method of communication in different ways which are by audio (spoken), text written and sign to exchange ideas, emotions, and information[2]. Nowadays, there are the vast amounts of information exchanging between one natural language to another in different formats using machine translation[1].

Wolaytta is a North Omotic language of the Ometo group spoken in the Wolaytta Zone and some other parts of the Southern Ethiopia such as Gamo, Gofa, Mello, Kucha and Dawro [2]. The number of Wolaytta language speakers were over 3.3 million and the sentence structure of the Wolaytta language is the same as the most spoken languages in Ethiopia which are Amharic and Afaan Oromo that means Subject Object Verb (SOV) word order [2][3]. The Wolaytta language uses Latin script written in the 1940’s[2]. Different type publications in Wolaytta language like books, literatures, and newspapers increasing over the last decade[2]. It is working language in the Wolaytta Zone of Ethiopia [1]. Currently, primary, secondary and higher institutions are using the Wolaytta language in teaching and learning processes in Wolaytta zone[2].

Machine translation is a technology for the automatic translation of text or speech from one natural language to another using computing device with or without human assistance[4][5]. In any translation, whether human or automated, the meaning of a text in the source language must be fully transferred to its equivalent meaning in the target language's translation. Translation is not a direct word to word substitution rather it considers grammatical meaning. It is one of the hottest research areas in computational linguistics that explores how computers can be utilized to understand and manipulate natural languages[1].

Advantages of machine translation are quick translation which is save the translation time translating sentences or paragraphs or documents within a short time, low price which is needs only language pair, not need the language expert or professional translator, confidentiality which is translate private or secret data, and universality[1]. Disadvantages of machine translation are ambiguity which means words in one language often map to multiple words in another language, idiomatic usage which means difficult to identify, phrase verbs are difficult to handle, structure difference among languages which means word order (subject-object-verb, subject-verb-object), and lexical differences[1].

For MT, a number of approaches are available. Such as Rule-Based MT (RBMT), Statistical Translation (SMT), Example-Based MT (EBMT), Hybrid MT and Neural MT (NMT) [5][6].

1.3. Statements of the Problem

Currently, the demand of translation is increasing rapidly but those demands are not fulfill using manually. Therefore, machine translation is rolling the great job by translating to satisfying the demand of translation. Machine translation is increases the benefit of web access by many non-native or non-English speakers by using freely open machine translation system which are Google Translate, Bing Translator, Yahoo! Babelfish and Systran[1].

According to Accredited[7][8] language websites the English language is the dominant language over the world. The English language is now the world's lingua franca and non-English speakers are faced with the problem of communication and limited access resources in English language. Most international scientific conferences and publications use English as their required language[6]. Therefore, the vast majority of scientific papers, most of the usable documents are published and available in English language[6].

Most of the works have been done from English to foreign languages to use these available documents, such as English-Spanish, English-Chinese, English-French, English-German, English-Russian, English-Portuguese, English-Japanese, English-Arabic, English to most of Indians languages and some of Ethiopian languages have been done which are English-Amharic, English-Afaan Oromo and English-Tigrigna have been attempted[9][10].

But, no one is attempt to develop English-Wolaytta language pair or vis verse machine translation. Therefore, the I believes that investigating the way of developing English-Wolaytta machine translation makes a direction to make available these documents in Wolaytta language.

Thus, the above mentioned reasons motivate me to investigating the development of English-Wolaytta machine translation system using statistical approach.

The research answers the following questions:

- What is the performance of English-Wolaytta SMT?
- Does the size of corpus has impact on the performance of the English-Wolaytta machine translation?
- Does morphological segmentation improve the quality of English-Wolaytta machine translation?

1.4. Objectives of the Study

1.4.1. General Objective

The general objective of this research is to develop a English-Wolaytta machine translation prototype using statistical approach.

1.4.2. Specific Objectives

To achieve the general objective of the study, the following specific objectives are identified.

- To review the related literatures from English to local language MT
- To review techniques and methodologies used for statistical MT
- To collect English-Wolaytta parallel corpus
- To collect monolingual corpora for the language modeling
- To develop English to Wolaytta prototype using statistical machine translation
- To evaluate the performance of the prototype

1.5. Methodologies

In order to achieve the objectives of this study, quantitative methodologies are applied. The following subsections discuss the methodologies that were followed in this study.

1.5.1 Literature Review

A detailed literature review has been done on machine translation on different language pairs in general and statistical approach of machine translation to explore the principles, methods, techniques and tools employed. Furthermore, related literatures, from English to foreign languages and syntactic relationship between English and Wolaytta languages has been reviewed.

1.5.2. Data Collection

To conduct statistical machine translation, parallel corpus of source and target language is required. The translation system tries to generate translations using the English-Wolaytta

corpus based on statistical methods. Since Wolaytta language has a very limited resources, the possible parallel English-Wolaytta sources are only the Holly Bible. Since then, collected 30,000 bilingual corpus (for each English and Wolaytta) from the Holly Bible only because of the scarcity of the Wolaytta language. Moreover, 38,993 monolingual corpus collected from educational documents and Holly Bible.

1.5.3. Tools and Techniques

Developed the English-Wolaytta machine translation system by using statistical machine translation approach, we have used most popular and freely available SMT tools such as: SRILM toolkit for language model, MGIZA++ align the corpus at word level by using IBM models (1-5). Decoding has been done using Moses, which a statistical machine translation system that is also used to train translation models to produce phrase tables. Ubuntu operating system which is suitable for Moses environment has been used. In addition, widely unsupervised morpheme segmentation tool Morfessor 2.0 is used for morphological segmentation of Wolaytta text.

1.5.4. Evaluation

Machine translation systems are evaluated by using automatic evaluation method or human 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.

In order to evaluate the performance of the prototype first we prepared the translated text by the system and second human translated text which used as reference translation selected randomly from Holly Bile, by using these two texts BLEU score metric evaluate the performance of the English-Wolaytta machine translation system.

1.6. Scope and Limitations of the Study

1.6.1. Scope of the Study

English-Wolaytta statistical machine translation is designed to translate a sentence written in English text into Wolaytta text.

1.6.2. Limitations of the Study

While conducting this study, some limitations have been faced. The major limitations are the lack of documents (parallel sentences), because of this; it leads to aligned 30,000 parallel sentences for both languages manually, and this by itself consumes time.

1.7. Contribution of the Study

Machine translation is playing great roles in exchanging information among different languages around the world. The rate of machine translation is faster than human translator.
In relation with this the main contribution of this research work are the following:

- Improving efficiency as compared to manual translation.
- The prepared parallel corpus uses for other natural language processing studies.
- The research done helps to adopt for other local languages, like Afaan Oromo to Wolaytta or Amharic to Wolaytta machine translation.

1.8. Organization of the Thesis

This thesis paper is organized into six chapters. The first chapter discusses about background, statement of the problem, objectives of the study, scope and limitation of the study, methodology followed including literature review, data collection, development tools and evaluation, and contribution of the study.

The second chapter deals with an overview of the Wolaytta language and its relationship and difference from English language. Beside this, it discuss in detail morphology analysis, writing system, sentence structure, punctuation marks and conjunctions of the Wolaytta language. The third chapter three deals with literature review of MT, approaches of MT such as RBMT, SMT, EBMT, HMT and NMT, morphology, alignment, automatic evaluation and related works.

Chapter four is the main part of this research, discussed about the design and development of the English-Wolaytta model including, corpus preparation, types of corpus used for the study, corpus alignment, and briefly discuss about the prototype of the system.

Chapter five briefly describes the process of the experiments or results of the system are discussed which included different experiments and the results of the experiments with interpretation of the findings. The last chapter is chapter six deals about the conclusion of the findings and recommendations for the further works.

CHAPTER TWO

2. WOLAYTTA LANGUAGE

2.1. Introduction

This chapter briefly discusses the different characteristics of Wolaytta words and sentences as compared to English language. The major Wolaytta word classes, which are nouns, verbs, adjectives and conjunctions, are described in this chapter.

2.2. Overview of Wolaytta Language

There are more than 89 languages in Ethiopia, Wolaytta language is one of them[1]. Some of Wolaytta people live out of their home land, Wolaytta is located in southern part of Ethiopia and 400 kilometers away from the capital city of Ethiopia (Addis Ababa). The name "Wolaytta" represents the nation, the land and the language.

As we mentioned in section 1.2, Wolaytta language is a North Omotic language of the Ometo group spoken in the Wolaytta Zone and around boundaries areas[2]. The native peoples call their language "Wolaytta" (Wolayttattuwa in their language). The language is also referred to as Wolaytta doonna (literally mouth of Wolaytta) or Wolaytta Kaalaa (literally word of Wolaytta). The English Bible translated into Wolaytta which finished in 2002[2]. The Wolaytta language is plays a crucial role for the people of the zone in social, economical, political and religious actives. It is used as an instructional media from primary schools to university level.

Furthermore, few literature works, newspapers, magazines, educational resources, official documents and religious writings are written and published in Wolaytta language[2][11][12]. Above 10,000 documents are written in Wolaytta language which are available in hard copy most of them are religious documents[2].

2.3. Morphology

Morphology is describes how words are formed in the language and it tries to discover the rules that govern the formation of words in a language [1][2][9]. The purpose of this stage of language processing is to break strings of language input into sets of tokens corresponding to discrete words, sub-words and punctuation forms[10][13]. For example a word like "unsuccessful" can be broken into three sub-word tokens as: un-success-ful. Morphology is

concerned primarily with recognizing how base words have been modified to form other words with similar meanings but often with different syntactic categories [9]. Modification typically occurs by the addition of prefixes and or postfixes cases of word form modification.

Morpheme is the smallest unit of meaning we have. That is, the smallest piece of a word that contributes meaning to a word, E.g. trainings (have three morphemes: train-ing-s)[1][9]. These are considered the basic units of meaning in particular language. Words that have meaning by themselves: boy, food, door, are lexical morphemes. Those words that function to specify the relationship between one lexical morpheme and other words like at, in, on, -ed, -s are called grammatical morphemes [9][10.

All morphemes are either free or bound. Free morpheme is one that can stand on its own, that is, it is an entire word, E.g. the, cat, run, pretty, trapezoid. Bound morpheme cannot stand on its own, but rather must be attached to a free morpheme whenever you say it. For example re-, un-, -est, -fer. Some morphemes are root the others are affixes. Affix is a morpheme which attaches to roots (stem), changing their meaning in regular ways. Affixes are generally either prefixes or suffixes. Prefix is an affix that goes before a root, such as re-read, and un-loved in English and eta-agaa, eta-asa, and eta-ayyo in Wolaytta. Suffix is an affix that goes after a root. E.g. -est, -er, -s (quick-est, quick-er, read-s, book-s). The affixes we just talk about are distinctive in one more way. They are acting in a particular way when attached to the base. Either they are giving grammatical information or they are creating a new word.

Inflection morphology is textual representations of words change because of their syntactic roles [1][13]. In English language the most popular nouns add -s (dog's) as a suffix, comparative and superlative forms of regular adjectives takes -er (talk-er) -est (long-est) suffixes. It is viewed as the process of adding very general meanings to existing words, not as the creation of new words.

Derivation morphology is derived new words from existing words. It is the process adding affixes combine with the roots to create new words, for examples 'global-globalism', 'happy-happiness', and read-readable'[1]. It is viewed as using words to make new words. Compounding is formed new words by grouping the existing words [1]. For example: headache, and toothpaste. The nature of morphological processing is heavily dependent on the language being analyzed.

There are two types of morphological approaches: morphological analysis and morphological synthesizer. Section 2.3.1 discussed about morphological analysis and section 2.3.2 discussed about morphological synthesizer.

2.3.1. Morphological Analysis

Morphological analysis was developed by Fritz Zwicky in the 1940's and 50's as a method for systematically structuring and investigating the total set of relationships contained in multidimensional, usually non-quantifiable, problem complexes[1][9]. Morphological analysis is to divide up (segment) whole words into their (smaller and smallest) constituent parts, with these parts themselves having meaning[9]. Morphological analysis is the identification of a stem form from a full word form[10]. The morphological analysis performs a recursive and exhaustive search on all possible segmentation of a given word and it takes a particular word as an input and it produces all possible segmentation of the word as an output[10]. Every segmentation should specify:

- A single stem in that word
- Each suffix in that word
- A syntactic analysis for the stem and each identified suffix

Once all the possible and correct segmentation of a word has been found, the morphological analysis combines the feature information from the stem and the suffixes encountered in the analyzed word to create a syntactic analysis that is returned [1][13].

2.3.2. Morphological Synthesis

Morphological synthesis is a process of returning one or more surface forms from a sequence of underlying (lexical) forms[1]. The morphological synthesis (generator) delivers a target language surface form for each target-language lexical form, by suitably inflecting it[9]. Morphological synthesis systems are used as components in many applications, including machine translation, spell-check, speech recognition, dictionary (lexicon) compilation, POS tagging, morphological analysis, conversational systems, automatic sentence construction and many others [1].

Morphological synthesis have vital role in natural language processing systems. They are used to generate surface word forms, which are the ones that are found in everyday communication, from lexical components that could be stored separately in different databases (lexicons). The combination of morphs to give meaningful words is the concern of morphological synthesis or generation. The morphological generator will synthesize the inflected word in its right form based on the morphological features [9].

2.4. Morphology of Wolaytta

As we discussed section 2.3, Morphology is a branch of linguistic and studies the internal structure of words[1][13]. Wolaytta is one of the morphological complex language and it uses both kinds of morphological inflection and derivation which morphological can result in very large numbers of variants for a single word [2]. The inflection morphology of Wolaytta language is 'siiqa', 'siiqaasa', 'siiqaasu', 'siiqada', 'siiqadasa', 'siiqadii', 'siiqaidda', 'siiqais', 'siiqoosona', and 'siiqida', it discussed more below. The derivation morphology of Wolaytta language is 'siiqakka', 'siiqasa' which derived from the root word 'siiqa'. It is the processes of forming a new word from the root or existing word often by adding suffixes (i.e. -kka, and -asa) in the language.

Affixation and compounding are two basic word formation processes in Wolaytta [1]. Affixation is a morphological process whereby a bound morpheme, an affix, is attached to a morphological base (root or stem). Affixes are morphemes that cannot occur independently. Prefix, infix and suffix are the three types of affixes [4]. A prefix occurs at the beginning of a word or stem (sub-mit, pre-determine, un-willing), a suffix at the end (wonder-ful, depend-ent, act-ion and some examples in Wolaytta siiqa-asa, kallo-kkona, and ciimma-tetta); and an infix occurs in the middle[11][16].

Morphemes are the smallest units of the meaning and that cannot be further decomposed into a meaningful unit [1][9]. For example, in English we have the word 'dog' which is 'kana' in Wolaytta language. 'Dog' (kana) is a complete idea it cannot be decomposed into smaller ideas based upon the word. There are two categories of morphemes: - free and bound morphemes. Free morpheme can stand as a word on its own whereas bound morpheme does not occur as a word its own. Both types of morpheme occur in Wolaytta language [14][15].

Among the Ethiopian languages which Amharic and Tigrigna uses both types of affixes, but the Wolaytta language does not have prefix and infix[2]. Instead, Suffixation is the basic way of word formation in Wolaytta which forming word by adding one suffix to another is common in Wolaytta. This process of adding one suffix to another suffix can result in relatively long word, which often contains an amount of semantic information equivalent to a whole English phrase, clause or sentence. Due to this complex morphological structure, a single Wolaytta word can give rise to a very large number of variants.

Compounding is a lexeme (less precisely, a word) that consists of more than stem [4]. It occurs when two or more words are joined to make one longer word, which function

independently. Although Wolaytta is very rich in compounds, compound morphemes are rare in Wolaytta and their formation process is irregular [2][16]. As a result, it is difficult to determine the stem of compounds from which the words are made.

2.4.1. Personal Pronouns

Personal pronouns are pronouns that are associated primarily with a particular grammatical person, first person (as I), second person (as you), third person (as he,as she). Personal pronouns may also take different forms depending on number (usually singular or plural), grammatical or natural gender, case, and formality.

In most languages basic distinctions based on personal pronouns. We see these distinctions within the basic set of independent personal pronouns [16]. The following table shows some examples.

English	**Wolaytta**
i	tanni
she	a
he	i
they	eti
we	nuuni
me	tana
us	nuna or nuussi
them	Eta

Table 2.1. Independent Personal Pronouns

2.4.2. Subject Verb Agreement

Wolaytta verbs often have additional morphology that indicates the person, number, and (second- and third-person singular) gender of the object of the verb [16].

ta michchiyo be7aas	
ta michchiyo	be7aas
my sister	I saw her
I saw my sister	

Table 2.2. Wolaytta Morphological Structure

2.4.3. Nouns

Wolaytta nouns subdivided in four main classes according to the endings they take in their inflection [1][16]. Thus, the classification of a noun as a member of the one or the other category rather depends on merely formal criteria. The nouns ending in -a in the absolute case and having stress on their last syllable belong to the 1st class. For example asa ‘person’ aawa ‘father’ tuma ‘darkness’. Nouns that have their absolute case ending in -iya and the stress on their penultimate make up the 2nd class. For example. morkkiya ‘enemy’, penggiya ‘the door’, siya ‘listen’. Nouns ending in -uwa in absolute case constitute the 3rd class. E.g. oduwa 'tale' metuwa 'trouble' fuuttuwa 'cotton'. The 4th class consists of nouns which end in - (i)yu in their absolute case. These mainly concern terms referring to female living beings. For example naiyu 'girl' bollotiyu 'mother-in-law' bakuliyu 'she-mule'.

2.4.4. Gender

The Wolaytta noun system, like most other languages, exhibits two different genders: "masculine" and "feminine" [2][16]. Accordingly, the nouns belonging to the 4th class are "feminine" while nouns belonging to all others (i.e. those of the 1st, 2nd and 3rd class) are defined "masculine". Formally, feminine differ from masculine by their endings.

The former end in –u in the absolute case where as masculine are all characterized by ending in –a (in the absolute case) and, if they are used as subject, additionally by the marker –y. Exceptions are possessive pronouns and some demonstratives. These may show forms which are limited to the one or to the other gender as shown below.

Masculine		**Feminine**	
dorssa	sheep	Dorssu	sheep
desha	goat	Deshu	goat
hagee	this	Hana	this
taagaa	he/it is mine	Taro	she/it is mine

Table 2.3. Wolaytta masculine and feminine

2.4.5. Number

The Wolaytta noun system comprehends two numbers, namely the singular and the plural [15] [16]. The singular usually consists of the basic noun form, while the plural is formed by means of a suffix used only for this purpose. Wolaytta forms the plural of nouns by means of a morpheme –tv, where –v represents a final vowel which changes according to the case inflection of the plural itself.

In the absolute case –v corresponds the –a, thus the plural marker is –ta in that case. Although all four noun classes form their plural by means of the morpheme –tv, the class membership of a noun also appears through the plural.

Thus, the nouns of the 1st class form their plural by means of the ending –a-ta [13].

Noun (Singular)		**Noun (Plural)**	
asa	person	Asata	Persons
qaala	word	Qaalata	Words
maxaafa	book	Maxaafata	Books
dorssa	sheep	Dorssata	Sheeps
paido	count	Paidota	Counts

Table 2.4. Wolaytta nouns singular and plural

Those of the 2nd class, instead, exhibit the ending –e-ta.

Noun (Singular)		**Noun (Plural)**	
Ogiya	Road	ogeta	Roads
Laaggiya	Friend	laggeta	Friends
Hariya	Donkey	hareta	Donkeys

Table 2.5. 2nd class Wolaytta nouns singular and plural

Those of the 3rd class are characterized by the ending –o-ta.

Noun (Singular)		**Noun (Plural)**	
Oduwa	tale	odota	Friends
Worduwa	lie	wordota	lies (lie peoples)

Table 2.6. 3rd Wolaytta nouns singular and plural

The nouns of the 4th class which consist of terms for female beings assume the plural form of their masculine counterpart.

Noun (Singular)		**Noun (Plural)**	
Imattiyu	female guest	imattata	female guests
Boogaanciyu	female robber	boogaancata	female robbers
Laagiyu	female friend	laageta	female friends

Table 2.7. 4th Wolaytta nouns singular and plural

2.5. Wolaytta Language Writing System

According the Demoze Wolaytta is an working language of literacy in Ethiopia and writing system was established in Latin script since 1970's [14]. Wolaytta language uses Latin based alphabet that consists of twenty nine basic letters of which five ('i', 'e', 'a', 'o' and 'u') are vowels, twenty four ('b', 'c', 'd', 'f', 'g', 'h', 'j', 'k', 'l', 'm', 'n', 'p', 'q', 'r', 's', 't', 'v', 'w', 'x', 'y', 'z' and '7') are consonants and seven pair letters fall together (a combination of two consonant characters such as 'ch','dh','ny', 'ph', 'sh', 'ts' and 'zh')[14]. The Latin alphabet is now adopted in mother tongue education system and various text books are published using Latin alphabet [16].

Latiine			Latiine		
Qeeraa	**Woggaa**	**Saaba**	**Qeera**	**Woggaa**	**Saaba**
a	A	ዐ	r	R	ሪ
b	B	ብ	s	S	ስ
c	C	ጭ	t	T	ት
d	D	ድ	u	U	ኡ
e	E	ኤ	v	V	ቭ
f	F	ፍ	w	W	ው
g	G	ግ	x	X	ጥ
h	H	ሀ	y	Y	ይ
i	I	እ	z	Z	ዝ
j	J	ጅ	ch	CH	ች
k	K	ክ	dh	DH	ደ
l	L	ል	ny	NY	ኝ
m	M	ም	ph	PH	ጵ
n	N	ን	sh	SH	ሽ
o	O	ኦ	ts	TS	ጽ
p	P	ፕ	zh	ZH	ዥ
q	Q	ቅ		7	ኧ

Table 2.8. Wolaytta language alphabet letter

2.6. Wolaytta Language Sentence Structure

Wolaytta and English languages have differences in their syntactic structure[16]. Wolaytta language uses Subject-Object-Verb whereas English use Subject-Verb-Object. In English sentence structure where the subject comes first, the verb second and the object third.

For instance, in Wolaytta language sentence, "Addaami kattaa immiis". "Addaami" is a subject, "kattaa" is an object and "immiis" is a verb. In case of English, the sentence structure is subject-verb-object. For example, if the above Wolaytta language sentence is translated into English it will be "Adam given food" where "Adam" is a subject, "given" is a verb and "food" is an object.

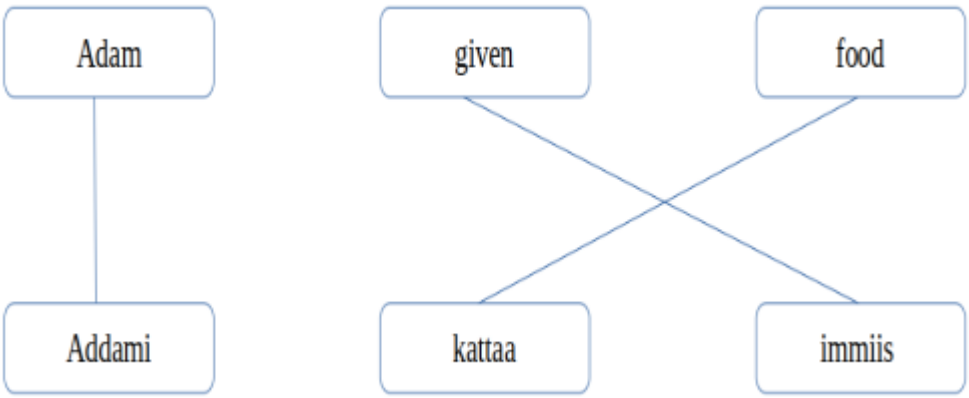

Figure 2.1. Wolaytta language sentence structure

2.7. Articles

In English there are three articles (a, an, and the) and it is used before nouns or noun equivalents and is a type of adjective [9]. The definite article which 'the' is used before a noun to indicate that the identity of the noun is known to the reader. It use before a singular or plural noun. The indefinite articles which 'a' and 'an' are used before a noun that is general or when it its identity is not known. The indefinite article which 'a' use before a singular noun beginning with a consonant and 'an' use before a singular noun beginning with a vowel[9].

Wolaytta language does not require articles that appear before nouns which the last vowel of the noun is dropped and added the suffixes (i.e. -ti) to show definiteness instead of using definite article [2]. For example, "people" to 'asaa' and 'the people' to 'asa-ti' dropped the last vowel (i.e. a) add the suffix -ti.

2.8. Punctuation Marks

Punctuation marks used in both Wolaytta language and English languages are the same and used for the same purpose with the exception of apostrophe. Some books using apostrophe on

glottal place 7 alphabet [19]. E.g. 'lo77o' using two apostrophe 'lo''o', 'de7iya' using one apostrophe 'de'iya'.

2.9. Conjunctions

Conjunction is a word used to connect words, phrases, clauses and sentences or to coordinate words [9]. In Wolaytta language there are different words that are used as conjunction. Some conjunction words in Wolaytta language are listed below.

English Conjunction	**Wolaytta Conjunction**
and	nne
or	woikko
so, therefore	hegaa gishshau
for	aissi giikko
but	shin
because	gishsha (do gishsha)
even if	hanikkokka
whenever	awudekka
wherever	awaanikka

Table 2.9. Wolaytta language conjunctions

CHAPTER THREE

3. LITERATURE REVIEW

3.1. Introduction

Under this chapter, there is a brief overview of machine translation and related works. The chapter briefly describes the machine translation which includes the approaches of machine translation those are Statistical Machine Translation (SMT), Rule Based Machine Translation (RBMT), Example Based Machine Translation (EBMT), Hybrid Machine Translation (HMT) and Neural Machine Translation (NMT).

3.2. Machine Translation (MT)

MT is an automated translation system and it is the process of translating text or speech from one natural language (source language) into another (target language) using computing device [1][9]. It is the part of computational linguistics and the hottest research area in computational linguistics that explores how computers can be utilized to understand and manipulate natural language. The idea of machine translation system is to produce the best possible translation with or without human assistance [1]. It is designed either for bilingual or multilingual system[9].

There are different types of important resources that are available on cloud (Internet) such as scientific, technical document, instruction manuals, legal documents, textbooks, bibles, newspaper reports, books, literatures, and most of them are published in some specific languages like English[10]. To use these published documents are difficult because of the written language differences. This is the situation where translations are needed to translate the resources. But it is very expensive and time consuming to translate manually among these languages [9][10]. Therefore, that is why machine translation is very important. Now a day, users around the world is using MT from different native languages for different purposes, like assimilation which is the translation of foreign material for the purpose of understanding the content, dissemination which is translating text for publication in other language and communication purpose like chatting and email. Currently, the most recent machine translations tools like Neural Machine Translation (NMT) achieving translation excellence [1].

Now a day, different types of machine translation approaches has been used to solve such language barriers. Machine translation for post-editing is the idea of improving the speed of

human translation by producing a draft translation [9]. As mentioned in the previous chapter, currently huge amount of information exchanging among natural language using machine translation. Machine translation is unable to directly produce publishable result, recent works in academia and industry has shown significant success [9].

There are different types of well-known machine translators that are available on-line openly and freely[1]. These are, Google Translate which was started by using statistical machine translation, but currently using neural machine translation. Nowadays, Google translating more than 100 languages and serves over 200 million people daily[1]. The second machine translator is Microsoft Bing Translator which is using statistical machine translation with having specific language rules translating over 60 languages [1]. The third machine translator is Systran translation.

Machine translation systems can be bilingual systems or multilingual systems depending on the number of languages involved in the process of translation [9][10]. Bilingual systems are designed specifically for two languages and multilingual systems are designed for more than two languages. In the case of unidirectional, the system translates from the source language into the target language only in one direction. Bidirectional systems work in both directions in a way that one language can act as source and the other as a target language and vice versa[9].

3.3. Approaches of Machine Translation

The machine translation system can be classified according to their methodology. There are two main approaches: the rule-based approach and the corpus-based approach [4][5]. In the rule-based (RBMT) approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts (linguistic professionals) is required. On the other hand, under the corpus-based (SMT and EBMT) approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Combining the features of the two classifications of machine translation systems give birth of the hybrid machine translation.

3.3.1. Statistical Machine Translation (SMT)

The issue of Statistical Machine Translation is introduced by Warren Weaver in 1949[9][10]. Statistical methods are applied to generate translation using large parallel corpus and it was began in 1988 by IBM researcher Peter Brown on the second TMI conference of the Carnegie Mellon University[4]. The basic principle of statistical machine translation is translation

decision based on probabilities of the corpus. Statistical methods are applied to generate translation using large parallel corpus to find the most probable target text sentence based on the parallel corpus for a given source text sentence [5]. The statistical machine translation system can be developed without the need of any language knowledge and it based only on bilingual sentence-aligned and monolingual data [5].

The translation model assigns a probability that a given source language sentence generates target language sentence [11]. The training corpus for the translation model is a sentence aligned parallel corpus of the languages. The translation accuracy of the SMT is highly dependent on its domain, size and quality of the parallel corpus.

Alignment

Alignment is the arrangement of something in an orderly manner in relation to something else[1][17]. It can be performed at different levels, sentences, word, morpheme level. Sentence alignment is aligning source language sentences from the target language sentences[1]. In this research the I aligned the parallel sentences manually because of the latency of parallel sentences.

Word alignment is the natural language processing task of identifying translation relationship among the words in a text, resulting in a bipartite graph between the two sides of the text, with an arc between two words if and only if they are translations of one another[17]. Iteratively redistribute probabilities until they identify most likely links for each word. Word alignment commonly has done using IBM Models 1-5. IBM model 1 is a straightforward application of EM, including the alignment to null token (deletion) finds translation probabilities for words in isolation, regardless of their position in parallel sentence [10]. IBM model 2-5 improve these distributions by considering: position of words in target sentence are related to position of words in source sentence, some source words may be translated into multiple target words (fertility of the words) and position of a target word may be related to position of neighboring words (relative distortion model)[9].

Language Modeling

Language modeling is to characterize, capture and exploit the restrictions imposed on the way in which words can be combined to form sentences [5]. It describes how words are arranged in a natural language. Language model is also gives the probability of a sentences (word order) or computing the probability of a sentence or sequence of words [1]. Language model

can be considered as computation of the probability of single word given all of the words that precede it in a sentence. It is an attempt to capture the inherent regularities (in word sequence) of a natural language. Language models are typically approximated by smoothed n-gram models and the probability is computed using n-gram model [9][10]. Statistical LMs estimate the probability of a sentence from its n-gram frequency counts in a monolingual corpus.

There are two approaches in language modeling:

- Grammar-based (defined based on linguistic knowledge), E.g. such as context free grammar, unification grammar
- Corpus based probabilistic approach (most widely used in NLP, and called statistical language modeling)

Monolingual corpus of the target language is defined as the language model. Language model is represents all sentences or word sequences which is a more operable abstraction of a language[9].

Translation Modeling

Statistical methods are applied to generate translation using bilingual corpora. This methodology uses different kinds of translation models. SMT search algorithm then determines the sentence by finding the highest product of the values sentence validity, word translation and word order [9]. SMT is the most widely used machine translation approach [9]. Statistical machine translation firstly started from a word-based translation but currently introduced other models phrase-based and syntax-based [5].

First SMT model word-based is counting of the probability of word by word translated into target language [5]. Word-based models translate words as atomic units. Word-based breaks down a sentence into smaller. There is difficulty in word-based in handling more than one word source language which has only one word translation in target language. And when one word source language can mean more than one word translation but not in sequential order will create a translation and handling syntactic transformation between the languages [9].

Second SMT models Phrasal-based is the sentence is cut into phrase segments. Phrase-based models translate phrases as atomic units and it is the most widely used model of statistical machine translation [10]. In phrase-based, the words are translated based on the phrases and once each phrase is translated, then they are reordered using the way of word alignment. Advantage of phrase-based translation are ability of translating many-to-many translation and use local context in translation[10]. If we provide more data, it will be very help full for more

phrases to be learned and mostly recommended [9]. Therefore, the I used phrase-based for this research.

Third SMT models Syntax-based statistical machine translation model is started with analyzing words of source language into its syntactic units rather than single words or strings of words[5][9]. Advantages of syntax-based SMT model is better reordering for syntactic rules such as following basic structure of position of subject, object and verb. And give a better explanation for function words such as preposition and determiners, as it analyzes each word in its syntactic position and the ability to put the syntactically related words in the right order [9].

Decoder

The task of decoding is finding the translation option that maximizes the log-liner mode is exponential in the input sentence [5]. The decoding starts by searching the phrase table for all possible translations for all possible fragments of the given sources sentences. Decoder is responsible for the search in the space of possible translations. Decoder uses feature scores and weights to select the most likely translation [1].

3.3.2. Rule Based Machine Translation (RBMT)

Rule based machine translation (RBMT) is the earliest machine translation approach. RBMT needs the morphological, syntactic and semantic information about both languages [17]. It is also known as "Knowledge-based Machine Translation" which is based on linguistic information about the source and target languages. Rules play major role in various stages of the translation, i.e. syntactic processing, semantic interpretation and contextual process of the languages[9][10]. There are three different types of approaches under the rule-based machine translation approach[9]. 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 [9].

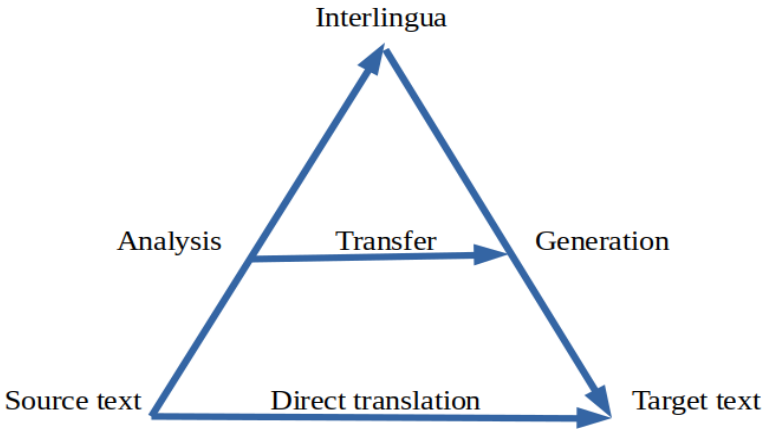

Figure 3.1. Different levels of analysis in an MT system

Direct approach is the first developed approach historically and it is adopted by most machine translation system and it is performed by considering individual words in the source language text translating each word to the target language text[1][10]. It uses a large bilingual dictionary, each of whose entries can be viewed as a small program with the job of translating one word. After the words are translated applied simple reordering rules. The advantages of direct translation are they are fast, simple, inexpensive and no translation rules hidden in lexicon. In the other hand, it has its own disadvantages, the disadvantages of direct translation are it misses any analysis of the internal structure of the source text, and lacks computational sophistication (poor translation quality).

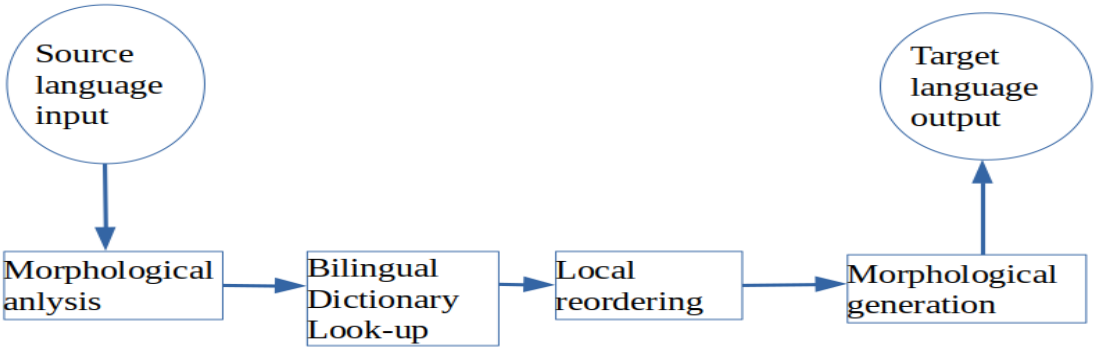

Figure 3.2. Direct MT system

Transfer translation is a set of rules ranging from morphology and syntax to semantics and context and it uses an intermediate representation that captures the structure of the original

text in order to generate the correct translation[1]. Transfer translation based approach fist parse the input text and then apply rules to transform the source language parse into a target language parse. The process of transfer-based translation involves: analysis, transfer and generation. Transfer bridges the gap between the output of the source language parser and the input to the target language generator[9][10]. Transfer based approach needs different types of rules: syntactic transfer, semantic transfer and lexical transfer.

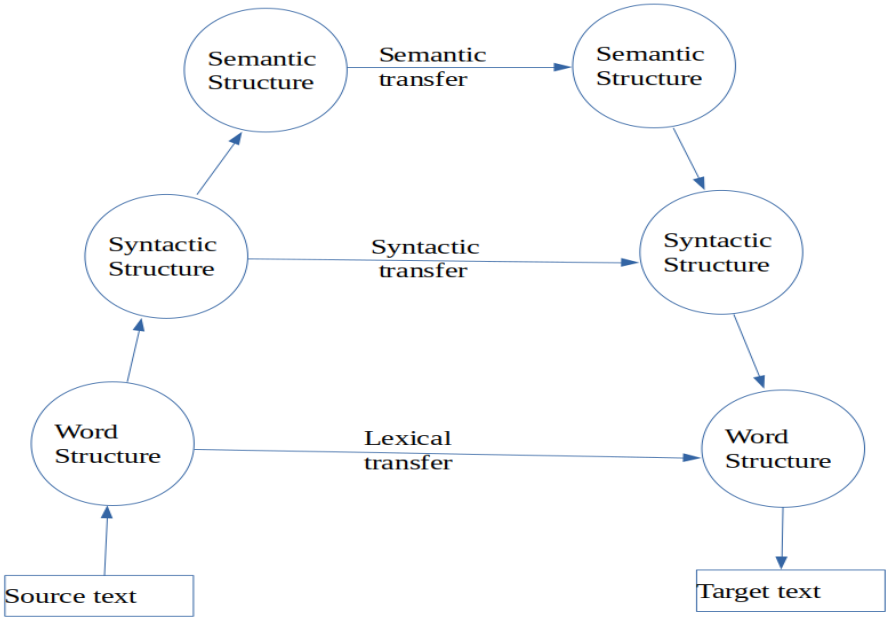

Figure 3.3. Transfer MT system

Transfer approach overcomes the language differences by adding structural and phrasal knowledge to the limitations of the direct approach [9]. Generally, English has SVO while Wolaytta has SOV structure. Transfer approach unifies this divergence by altering the structure of the input sentence to make it conform to the rules of the target language. There are different types of advantages of transfer based approach, such as offers the ability to deal with more complex source language phenomena than the direct approach, high quality translations can be achieved as compared to direct approach and relatively fast as compared to interlingua translation[9].

Interlingua translation is based on the utopia of a neural language that would be able to represent all meaningful information of every utterance in every language [5]. The idea is for the interlingua to represent all sentences that mean the same thing in the same way, regardless of the language. In inter lingua approach the source language text is analyzed into an abstract meaning representation called an Interlingua[1][5]. Then the target language text is generated from this representation or on other meaning his model translates by performing deep

semantic analysis on the input from language A into the Interlingua representation and generating from the Interlingua to language B. This approach is appropriate for many-to-many translations as it reduces the language pair interdependence, each language is dependent on the meaning (Interlingua) and not on another language. The process of Interlingua translation involves: analysis and generation. The aim of analysis is the derivation of an Interlingua representation [1][4].

There are different types of advantages of Interlingua, such as the most attractive for multilingual systems (the addition of a new language entails the creation of two modules), permits translation from and into the same language (valuable during system development in order to test analysis and generation modules)[5].

Vauquois triangle is a common way to visually present the above three approaches. The vauquois triangle shows the increasing depth of analysis required on both the analysis and generation end as we move from the direct approach through transfer approaches to Interlingua approaches [4]. It also shows the decreasing amount of transfer knowledge needed as we move up the triangle, i.e. From huge amounts of transfer at the direct level (almost all knowledge is transfer knowledge for each word) through transfer (transfer rules only for parse tree or thematic roles) through Interlingua (no specific transfer knowledge).

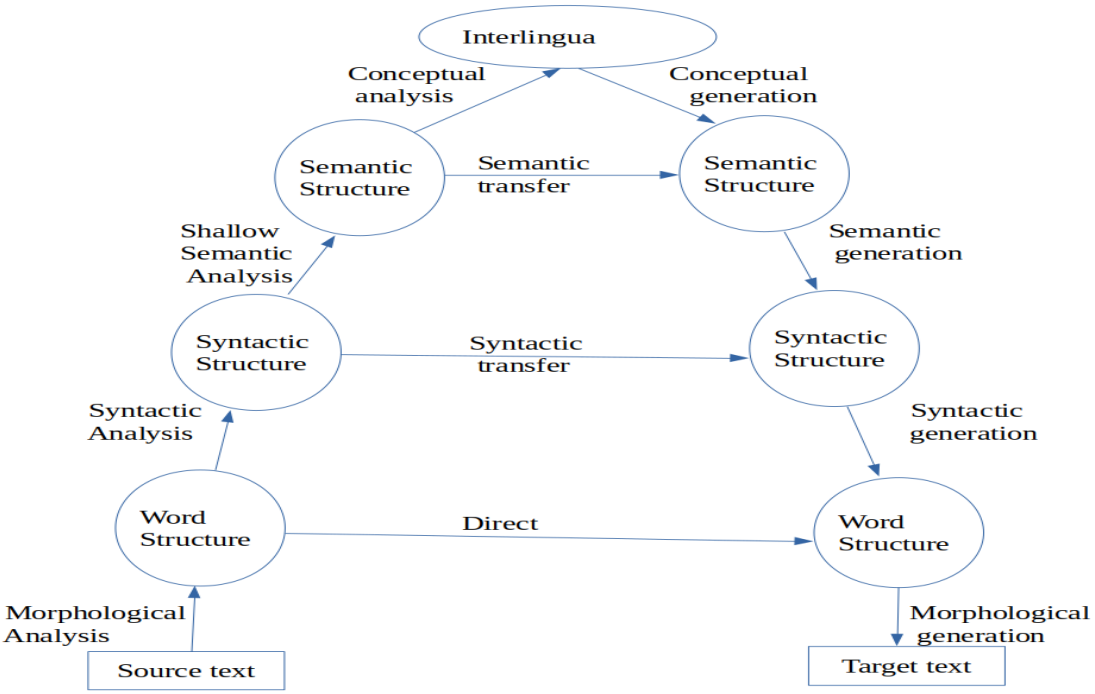

Figure 3.4. Vauquois MT system

3.3.3. Example Based Machine Translation (EBMT)

EBMT is a set of phrases in the source language and their corresponding translations in the target language are given the example database [9][10]. EBMT is based on finding (recalling) the analogous examples of the language pairs. The EBMT system uses these examples to translate new similar sources language phrases into the target language. The basic premise is that if a previously translated phrase occurs again the same translation is likely to be correct again. The idea behind EBMT is people do not translate by doing deep linguistic analysis of a sentence. The critical issue of the EBMT is needs a large scale bilingual corpus, which can be unlimited in for good performances. For the best translation quality, the training corpus should be huge corpus.

There are three steps in EBMT, such as matching the source language input against the example database, alignment or adaption selecting the corresponding fragment in the target language and then recombination (target sentence generation or synthesis) and recombining the target language fragments to form a correct text [1][4]. For example, linguistic knowledge about word order, agreement, etc. are captured automatically from examples. It relies on large corpora and tries somewhat to reject traditional linguistic notions like part of speech and morphology [4][9]. EBMT systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs.

The basic idea is to collect a bilingual corpus of translation pairs and then use a best match algorithm (using the distance of the input string from each example translations) to find the closest example to the input string in question. The translation quality of EBMT is based on the size of quality corpus and the quality increases as the examples become more and more complete. EBMT is also efficient as (in its best case) it does not involve application of complex rules and rather finds the best example that matches to the input (using some distance calculations) [4]. To see how EBMT works, consider the example of translating the sentence,

Example-based machine translation (EBMT database)	
English language sentence	Wolaytta language sentence
E.g. The price of the book is more than 500 Birr.	E.g. Maxaafaa miishshaayi 500 biraappe daro.
E.g. The price of the house is very good	E.g. Keetta miishshaayi daro lo77o
Based on the above EBMT examples translations the following sentence is translated	
E.g. The price of house is more than 500 Birr.	Keetta miishshaayi 500 biraappe daro.

EMBT based on the above examples takes the first phrase from the 2nd sentence (The prices of the house) and takes the second phrase from the 1st sentence (is more than 500 birr).
The advantage of EBMT: uses fragments of human translation (previously translated texts or sentence from example based database) which can result in higher quality. The disadvantage of EBMT: may have limited coverage depending on the size of the example database.

3.3.4. Hybrid Machine Translation (HMT)

Hybrid Machine Translation is a method of MT that is characterized by the use of multiple machine translation approaches within a single machine translation system [10]. The motivation for developing hybrid machine translation systems from the failure of any single technique is to achieve a satisfactory level of accuracy. There are many hybrid machine translation systems have been successful in improving the accuracy of the translations and there are several popular machine translation systems which employ hybrid methods, i.e. PROMT, SYSTRAN and Omniscib Technologies (Asia Online) [1][4].

Hybrid systems have been developed by combining the positive sides of each of the above approaches (SMT and RBMT). Because of each machine translation approach has its advantages and disadvantages. Hybrid systems take the synergy effect of SMT and RBMT[10]. Hybrid machine translation approach leverages the strengths of the approaches (SMT and RBMT). In HMT architecture there are three components of HMT architecture: identification of source language by observing chunks (words, phrases and equivalents), transformation of the chunks into target language, and generation of translated language [10].

3.3.5. Neural Machine Translation (NMT)

The current state-of-the-art is Neural Machine Translation (NMT). NMT is an end-to-end learning approach for automated translation with the potential to overcome many of the weakness of conventional phrase-based translation systems [1] [18]. Instead, people have been turning their heads towards neural machine translation (NMT) systems, which after being introduced seriously in 2014 have seen many refinements. These systems are also

known as sequence-to-sequence models or encoder-decoder networks, and were initially fairly simple neural network models made out of two recurrent parts [18].

First, an encoder part taking an input sentence in the source language and computing an internal representation, and secondly the decoder part, a neural language model trained to be good at assigning a high probability to a well-formed sentence in the target language, which can be used to generate sentences that sound very good [18]. Letting the language model be conditioned on the input turns it into a translation model.

These early NMT systems worked on word level, which means that a word is seen as a symbol, so the input to the encoder is a unique index for each unique word, and the decoder being constrained to pick words from this vocabulary [1][18]. These models worked well and got some promising scores in evaluations, but they had some drawbacks. Firstly, the longer the input sentence, the more difficult for the encoder to capture all important information in the internal fixed-size representation. Secondly, they are practical only for use with a fairly limited vocabulary size. Neural MT is the approach of modeling the entire MT process via one big artificial neural network (ANN) and it is directly maps a source sentence into a target sentence within a probabilistic framework [18].

There are three big wins of Neural MT are end-to-end training which is all parameters are simultaneously optimized to minimize a loss function on the network's output, distributed representation share strength which is better exploitation of the word and phrase similarities, better exploitation of context which is NMT can use a much bigger context both source and partial target text to translate more accurately [18].

The main advantages of neural machine translation(NMT) are no need domain knowledge, no need to store explicit translation model and language model, can jointly train multiple features, and can implement decoder easily[1]. The main disadvantages of neural machine translation (NMT) are time consuming to train NMT model (two weeks, but depend based on the corpus size), slow in decoding (takes a weeks), if target vocabulary is large, weak to OOV (out of vocabulary) problem, difficult to debug the errors, and needs high perform computing devices (GPU - graphic process unit) [1].

3.4. Evaluation of Machine Translation

Being able to evaluate a machine translation system is crucial. Evaluating the quality of machine translation is an extremely subjective task. Translations are evaluated along fidelity, fluency and combination of the two (fidelity and fluency). Evaluation of machine translation

has proven to be quite difficult. The most accurate evaluations use human raters to evaluate each translation along fidelity and fluency [17].

Along fluency:

- How intelligible, how clear, how readable, how natural the MT output .
- One method is to give the raters a scale and ask them to rate each sentence of the MT output.
- Other methods rely less on conscious decision of the raters (we may measure the time it takes for the rates to read each output sentence).

Along fidelity:

- We measure adequacy and in informativeness
- Adequacy is judged by whether it contains the information that existed in the original
- The in informativeness of a translation is a task based evaluation of whether the information in the MT output is sufficient to perform some task

Combination of fidelity and fluency:

- Edit cost of post-editing: we can measure the number of words, the amount of time, or the number of keystrokes required for a human to correct the MT output to an acceptable level

There are different type of automatic evaluation metrics to evaluate machine translation system. Three major metrics: BLEU, METEOR and WER,

Bilingual Evaluation Understudy (BLEU) is evaluate machine translation output matching from human reference translation [19]. Therefore, a test corpus is needed for this method, given at least one manual translation for each test sentence. During a test, each test sentence is passed to the MT system, and the output is scored by comparison with the correct translations. This score is called the BLEU score[19].

The BLEU score is evaluated by two factors, concerning the precision and the length of candidates, respectively. Precision refers to the percentage of correct n-grams in the candidate. In the simplest case, unigram (n=1) precision equals to the number of words from the candidate that appear in the references divided by the total number of words in the candidate [9][19]. The BLEU is done based on the modified n-gram precision, which is calculated by dividing the number of n-grams in the translation that matches an n-gram in a reference, by the total number of n-grams in the translation. This is called modified, since each reference n-gram is only allowed to match once. BLEU is the most commonly (widely) used metric[1][9].

Metric for Evaluation of Translation with Explicit Ordering (METEOR) is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision[1] [9]. It was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgment at the sentence or segment level[1]. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

Word Error Rate (WER) is a common metric of the performance of a speech recognition or machine translation system and it works at the word level[1]. It was originally used for measuring the performance of speech recognition systems, but is also used in the evaluation of machine translation[1]. The metric is based on the calculation of the number of words that differ between a piece of machine translated text and a reference translation[1].

3.5. Related Works

There are different studies conducted on machine translation approaches, strategies, techniques and implementations has been documented. In Ethiopia, some machine translation systems has been tried to be developed and documented as a research work. Out of these, some are explained in brief under the following subtopics.

3.5.1. English–Afaan Oromo Machine Translation: An Experiment Using Statistical Approach

The first attempt for statistical machine translation for local conducted [6]. In general, the study had two main goals: the first is to test how far one can go with the available limited parallel corpus for English-Afaan Oromo language pair and the applicability of existing statistical machine translation systems on these language pairs. The second one is to analyze the output of the system with the objective of identifying the challenges that need to be addressed.

The architecture includes four basic components of statistical machine translation, which are language modeling, translation modeling, decoding and evaluation. The language modeling component takes the monolingual corpus and produces the language model for the Afaan Oromo (target) language. The translation model component takes the part of the bilingual corpus (English and Afaan Oromo) as input and produces the translation model for the given language pairs. The decoding component takes the language model, translation model and the source text to search and produce the best translation of the given text. The evaluation

component of the system takes system output and the reference translation of the input to produce the metric that compares the system output and the reference translation.

Collected parallel sentences from different resources for this study, such as Constitution of FDRE (Federal Democratic Republic of Ethiopia), Universal Declaration of Human Right, proclamations of the Council of Oromia region State, religious documents and already translated and available documents. The documents were preprocessed by using different scripts which are customized to handle some special behaviors of the language 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. Used SRILM toolkit for language modeling, MGIZA++ for word alignment and Moses decoder was used for decoding purpose. By using these resources and toolkits, an average BLEU score of 17.74% was achieved based on the experimentation.

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

The study which was done with the objective of developing a bidirectional English-Amharic machine translation system using constrained corpus[9]. The study was implemented by using statistical machine translation approach. Prepared two different corpora: the first corpus (Corpus I) was made of 1,020 simple sentences that has been prepared manually. Out of the 1,020 simple sentences, all the sentences were used for the training set and 10% of the corpus 102 simple sentences for testing set.

The second is made of 1,951 complex sentences (Corpus II) from sources, one from Bible and the other from the public procurement directive of Ministry of Finance and Economic Development. The Corpus II is like the experiment on Corpus I, the experimentation process consisted of two methodologies. Out of the 1,951 complex sentences 2% were taken for the testing process which is around 40 sentences. This is because of the complex sentence is very large and complicated which makes it hard for the candidates to assess the translation.

Since the translation is bidirectional, two language models were developed, one for Amharic and the other for English and translation models were also built. A decoder which searches for the shortest path was used and expectation maximization algorithm was used for aligning words in the accurate order.

Two different experiments were conducted and the evaluation was performed by using two different methodologies. The first experiment was performed using simple sentences and

evaluated by using manual using BLEU score has an average score of 82.22% accuracy for English to Amharic translation and 90.59% for Amharic to English translation. The second experiment was performed by using the manual questionnaire method, accuracy of English to Amharic translation is 91% and accuracy of Amharic to English translation is 97%.

For complex sentences, the first methodology, the accuracy of the translation from English to Amharic was 73.38% and from Amharic to English translation was 84.12%. The second experiment was performed by using the manual questionnaire method was 87% for the English to Amharic translation and 89% for Amharic to English translation. The study shows Amharic to English translation has a better accuracy than English to Amharic translation.

3.5.3. Preliminary Experiments on English-Amharic Statistical Machine Translation (EASMT)

A preliminary experiments on English-Amharic statistical machine translation is conducted[20]. Used English-Amharic parallel corpus from parliamentary documents that exist on-line including those collected manually are used for the preliminary experiment on EASMT. Conducted preprocessing tasks was performed on the parallel sentences in order to retain and convert the full content into a valid format suitable for the EASMT system. Some of the preprocesses included text conversion, trimming, sentence splitting, sentence aligning and tokenization. The process of trimming is performed before and after aligning at document level. The sentence splitting has been done before starting aligning at sentence level while tokenization is performed after aligning at the sentence level. The alignment at the sentence level has been done using a sentence aligner called Hunalign.

The Amharic sentence endings and punctuations have been converted to English to make it easy to apply similar preprocessing tools used for English. The converted Amharic punctuations include the Ethiopian comma(፣), colon(፡), semi-colon(፤) and full stop(።) to their English counter parts comma(,), colon(:), semi-colon(;) and full stop(.). 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.

Out of the total collected sentences, 90% (16,432) randomly selected sentence pairs have been used for training while the reaming 10% (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 language modeling which collected from Ethiopian News Agency not included bilingual corpora.

The proposed architecture includes the basic components of SMT, such as LM, TM, decoder and evaluation metrics. The used toolkits are to implement the EASMT system: SRILM for LM, MGIZA++ for word alignment, Moses for decoding and Per scripts for pre-processing purposes.

Developed two types of SMT which are segmented and unsegmented. The unsegmented is the baseline system which is normally trained, developed and evaluated using the corpus. The segmented system has been developed by segmenting all the Amharic texts. The BLEU score of the segmented system is 36.58%, which is a 0.92% increase from the baseline (unsegmented) system that has been a BLEU score of 35.66% by applying morpheme segmentation.

3.5.4. Bidirectional English–Afaan Oromo Machine Translation Using Hybrid Approach

Moreover, an attempted with the intention of using bidirectional machine translation[9]. The research work is implemented using hybrid of rule based and statistical approaches. Collected corpus from different domains and sources including Holly Bible, Constitution of FDRE, and Criminal Code of FDRE, international conventions, Megeleta Oromia and a bulletin from Oromia health bureau from websites and offices. And prepared in a format suitable for use in the development process and classified as training set and testing set. After prepared corpus performed tokenization, true-casing and cleaning to make ready for experiment.

Collected 3,000 parallel sentences. Out of the those sentences used 90% of it which is 2,900 sentences are used for training and the remaining 10% which is 1,000 sentences used for testing. Since the system is bidirectional, two language models are developed, such as one for English and the other for Afaan Oromo. Translation models which assign a probability that a given source language text generates a target language text are built and a decoder which searches for the shortest path is used. Used different freely open source available toolkits to develop the system, such as IRSTIL toolkit for language modeling, MGIZA++ for word alignment, and Moses for decoding purpose.

The study was carried out with two experiments which are conducted by using two different approaches and their results are recorded. The first experiment is carried out by using a statistical approach. The result obtained from the experiment has a BLEU score of 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation. The second experiment is carried out by using a hybrid approach and the result obtained has a BLEU score of 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo

to English translation. From the result, we can see that the hybrid approach is better than the statistical approach for the language pair and a better translation is acquired when Afaan Oromo is used as a source language and English is used as a target language.

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

3.5.5. English-Tigrigna Factored Statistical Machine Translation

The study which was done[21]. English-Tigrigna translation was conducted using statistical machine translation approach. Collected 17,649 parallel sentences from Bible. Out of the 17,000 sentences used for training set and the remain 649 sentences for testing set. For monolingual corpus additional sentences from the international news website has been used.

The experiment of the system on three types which are baseline, segmented and factored corpus. The baseline setup is plain phrase-based translation model (single factored). The segmented translation system was trained and evaluated using factored corpus. While doing pre-processed sentence level segmentation and tokenization, a program codes which is written Python has been used. After preprocessed morphological segmentation, stemmed and POS tagged were performed to prepare the factored corpora. Used different type of toolkits for this study which are Python for preprocessing activities, HUNALIGN for sentence alignment, MGIZA++ for word alignment, SRILM for language modeling, Moses decoder for automate the system (statistical machine translation), Morfessor for morphological segmentation, and BLEU for evaluate the system.

The performance of the system was tested using BLEU metric. The performance of the baseline system BLEU score was 21.04%, segmented BLEU score was 22.65% which is increased by 1.61% from the baseline system and factored corpus BLEU score was 16.51 which is decreased by 4.53% from the baseline system and by 6.15% from segmented system. See that the segmentation has contributed for the overall performance of the segmented system that has shown better performance compared to the baseline system. The factored corpus has shown a decrease of 6.15% from the segmented and 4.53% from the baseline system. The low performance of the factored system is accounted to the POS tags attached since the tagger was trained using a small manually tagged corpus.

3.5.6. Bidirectional Tigrigna-English Statistical Machine Translation

This study was done[22]. The bidirectional English-Tigrigna translation was conducted using statistical machine translation approach. Collected corpora from different domain and classified into five sets of corpora, such as the first (Corpus I) from New Testament contains around 6,106 parallel sentences. The second (Corpus II) customized simple sentences from Eline[10], which contains around 1,160 sentences manually aligned at sentence level. The third (Corpus III) combination of the Corpus II (customized simple sentences) and FDRE sentences, which contains around 1,982 sentences. The fourth (Corpus V) corpus is prepared by combining Corpus I; sentences from the Holly Bible, and Corpus III; sentences from constitution of FDRE and customized simple sentences which contains around 8,088 sentences. The fifth (Corpus V) corpus is prepared from Corpus I and Corpus III by considering their size which contains around 3,000 sentences and prepared all corpus in a format suitable for use in the development process.

Prepared two monolingual corpora; one for English (16,016 sentences) and the other for Tigrigna (8,506 sentences). In addition to the sentences extracted from the bilingual corpora, additional sentences are added from different sources; with the objective of increasing the size of the monolingual corpus that is used to estimate the fluency of a sentence in the language. Additional sentences for English monolingual corpora are from criminal code of Ethiopia, and English version of Tigrigna regional state constitution. For Tigrigna monolingual corpora, additional sentences are added from the web. Preprocessed included sentence alignment, tokenization, and true-casing. Used different type of freely available software, such as IRSTLM for language modeling, MGIZA++ for word alignment, Moses used to train the system in both directions, Morfessor 1.0 for morpheme segmentation and BLEU score for evaluation of the system.

The experiments divided in three sets, such as baseline (phrase-based machine translation), morph-based (based on morphemes obtained using unsupervised method) and post processed segmented systems (based on morphemes obtained by post-processing the output of the unsupervised segmenter). Used 90% for training and the reaming 10% for testing on both directions in each corpus sets. Accordingly, the result obtained from the post processed experiment using Corpus II has outperformed the other, and the result obtained has a BLEU score of 53.35% for Tigrigna-English and 22.46% for English-Tigrigna. The research clearly shows that the post segmented system outperforms all the other experiments. Therefore, recommended future research should focus to further improve the BLEU score by applying preprocessing and post-processing techniques.

CHAPTER FOUR

4. DEVELOPMENT OF ENGLISH-WOLAYTTA SMT

4.1. Introduction

Under this chapter, the English-Wolaytta machine translation using statistical approach will be discussed in detail. Accordingly, the overall architecture of the system and its components has been discussed.

4.2. Architecture of the English-Wolaytta SMT

The following figure 4.1. Shows the architecture of the English-Wolaytta statistical machine translation and explain the meaning of each parts in detail below.

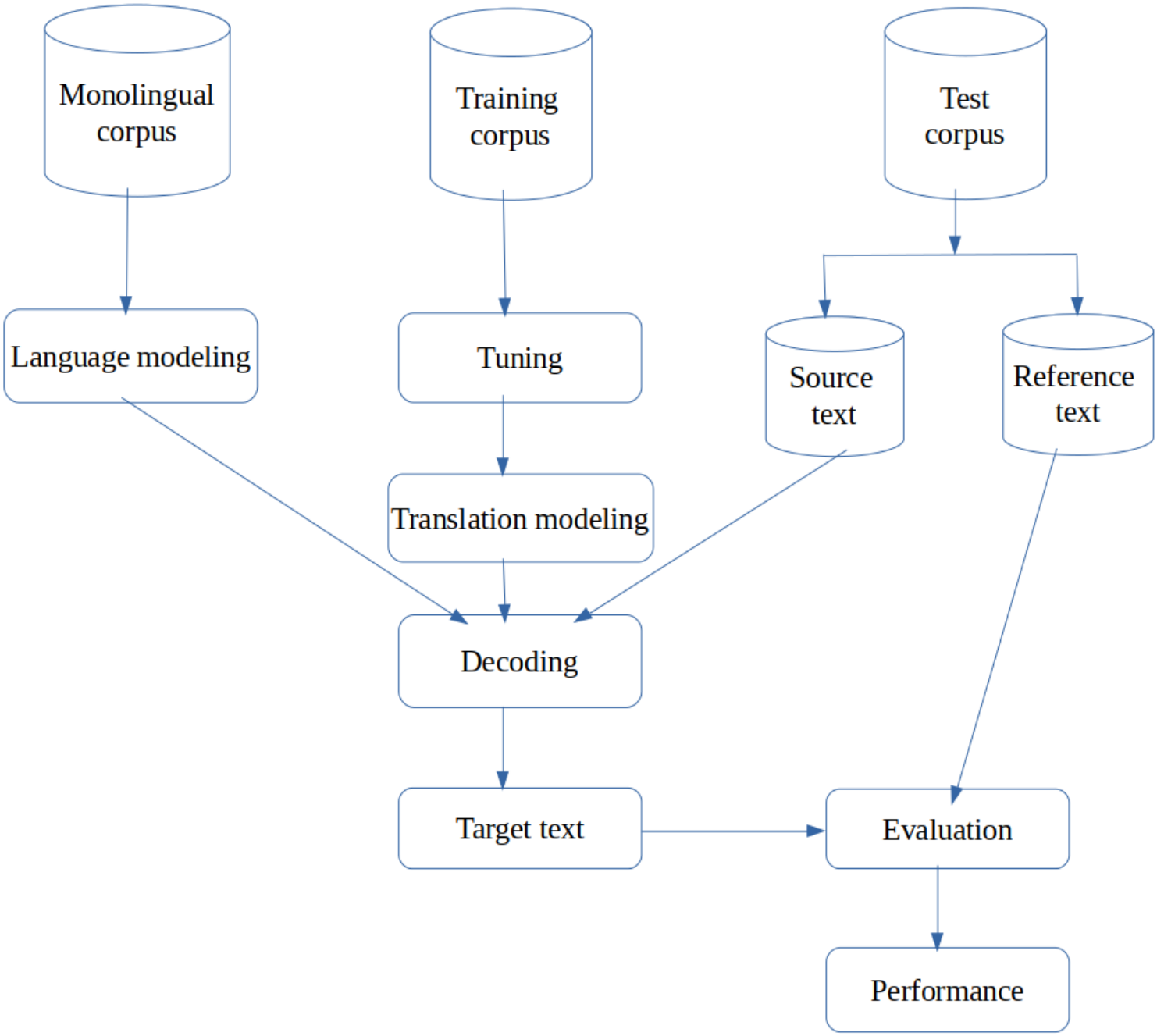

Figure 4.1. Architecture of SMT

As shown on the above figure 4.1, Monolingual corpus uses to develop the language model for the target language. The training corpus prepared from both languages, which was used to develop English-Wolaytta translation model. After we use translation model, then tuning has been implemented to maximize the translation performance. Test corpus prepared from both languages which was source text and reference text. Decoder is used to predict words in the target language using the language model, translation model (after tuned) and source text and

then produce the target text. Evaluation is conducted by comparing the output of translation system and reference text then produced the performance of the translation system. The above architecture again used after morphological segmentation (used Morfessor software) for the Wolaytta texts which are to improve the performance of the translation system. Each component is discussed in the following subtopics.

4.3. Corpus Collection and Preparation

In Wolaytta language getting in electronic format was not an easy task. Even though there are more than 10,000 hard copy documents are translated from other languages to Wolaytta language, but these documents are still not available in electronic format (scarce) and not easily accessible[2]. Tried to get digital format from different sources, such as from Wolaytta Sodo University department of language, Wolaytta zone education office for monolingual corpus and bilingual corpus from the Holly Bible.

4.3.1. Preliminary Preparation

The collected files are in different formats. Therefore, manipulate the data to put into uniform formats and encoding was necessary. Some of the documents are doc (Microsoft office) files, PDF and HTML files. One of the major difficulty in the preparation task was prepare parallel pair (corpus) manually for both languages because of the scarcity of parallel corpus. All of the data in the collected corpus was subsequently converted to plain text, cleaned up from the blank lines and noisy characters, and its encoding was converted to UTF-8 automatically to make it ready for training of the system. There is a number of preprocessing to get a cleaned corpus. The corpus was prepared in a format that needs to be applicable for the system. These preprocessing includes sentence alignment, normalization, and tokenization, lower-case and cleaning.

Sentence alignment is aligning each sentence which is aligned 30,000 English sentences with Wolaytta sentences pairs manually because of the scarcity of parallel corpus. Tokenization is the process of separating words from the punctuation marks and symbols into tokens [1]. These tokens are words, punctuation marks, and numbers. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. The tokenization tasks are performed on the parallel corpus in order to convert them into suitable tokens for the translation system using the tokenization script. In normalization, which is the replacing the following special characters such as "and" to ", ‘and’ to ' using the normalization script. Lower-case this is changing capital letters to small letters using the lower-case script. In Wolaytta language their is no different meaning between lower and upper case words. Cleaning is removing a long and empty sentence which

causes problems while training the pipeline, and obviously misaligned sentences are removed using the cleaning script. The all mentioned preprocesses (tokenization, normalization, punctuation and cleaning) are suitabled the parallel corpus for train translation system .

4.3.2. Bilingual Corpus

The corpus contains 30,000 sentences only because of the scarce of electronic Wolaytta language document which is very small size sentences comparing the most documents available languages; therefore we can not compare the Wolaytta language from other available resources languages like Indo-European language pairs, and EUROPARL corpus.

Units	**Bilingual**	
	English	Wolaytta
Sentences	30000	30000

Table 4.1. Bilingual sentences

4.3.3. Monolingual Corpus

The monolingual corpus is used for language modeling (target language) and it is necessary for training the fluency of the target language. Prepared the monolingual corpus from Bible (excluded test set), educational documents from Wolaytta Sodo University and Wolaytta zone education office texts. The monolingual corpus contains 38,993 sentences.

No.	**Monolingual sentences (Wolaytta language)**
1	28,500 sentences from training set
2	600 sentences from tuning set
3	9,893 sentences from other documents
	Total 38,993 sentences

Table 4.2. Monolingual sentences

4.3.4. Language Model

Language model is the probability of a sentence or sequence of words which will be trained from monolingual corpus for the target language. There are various software packages available to build statistical language model. SRILM is one of them and in this study; this software is used [22]. SRILM toolkit is composed of set of tools for building and applying statistical language models. It has been under development in the SRI (Speech Technology and Research Laboratory) since 1995[1][4]. The main purpose of SRILM is to support

language model estimation. For this study, uses the SRILM tool n-gram count to estimate two language models. The first language model is built upon the baseline (unsegmented) and the second (segmented) language model is after morphological segmentation has done for the Wolaytta language.

The probabilities obtained from the n-gram model could be unigram, bigram, trigram or higher order n-grams.

For example, given the following Wolaytta sentences,

- simooni qumaa miis
- simooni qumaa mibeenna
- simooni dabbotukko biis
- martta moliyaa maasu
- hewaana soo baasu
- martta Hewaani michchiyo

The unigram probability can be computed by:

$$P(\mathbf{w_1}) = \frac{count(w_1)}{\text{total words observed}} \Rightarrow P(\mathbf{simooni}) = \frac{3}{18} = \mathbf{0.167}$$

Where 3 are the number of times the word 'simooni' was used and 18 is the total words in the corpus (sample corpus).

The bigram probability can be computed by:

$$P(\mathbf{w_2}|\mathbf{w_1}) = \frac{count(w_1 w_2)}{count(w_1)}$$

$$\Rightarrow P(\mathbf{qumaa}|\mathbf{simooni}) = \frac{count(simooni\ qumma)}{count(simooni)} = \frac{2}{3} = \mathbf{0.667}$$

Where 2 is the number of times the words 'simooni' and 'qumma' have been happened together in corpus and 3 is the number of times the word 'simooni' is happen in corpus

And the trigram probability becomes:

$$P(\mathbf{w_3}|\mathbf{w_1}\mathbf{w_2}) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

$$\Rightarrow P(\mathbf{miis}|\mathbf{simooni\ qumma}) = \frac{\text{count(simooni qumma miis)}}{count(simooniqumma)} = \frac{1}{2} = \mathbf{0.5}$$

Where 1 is the number of times 'simooni', 'qumma' and 'miis' have been happened together and the 2 is number of time 'simooni' and 'qumma' have been happened together. Language modeling is the probability of fluent or grammatical sentence estimated using n-gram model from monolingual corpus. There are different languages modeling softwares that are available freely on the Internet. For this study, used SRILM toolkit to develop language model [22].

For the corpus with sentences, the n-gram model performs well with the unigram, bigram and trigram models. Trigram model was used based on the nature of the corpus that is used for the language model. 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 should not be approximated to zero. .

4.3.5. Translation Model

Statistical methods are applied to generate translation model using bilingual corpus. Based on the above architecture, two bilingual corpus are prepared for the translation model. The first bilingual corpus is for the baseline (unsegmented) translation and the second is after morphological segmentation (segmented bilingual corpus). The morphological segmentation only for Wolaytta text. The translation model assigns the probability of a given source language which will generate the target language sentence. The role of the translation model is to find the highest probability (P) of the source (e) text given the translated sentence to target (w) text. In this case the training corpus for the translation model is a sentence-aligned parallel corpus of the English languages to Wolaytta language. To find the approximate of the sentence translation probability the translation probabilities of chunks in the sentences is used, the chunk translation probabilities can be found from the bilingual corpus using expectation maximization algorithm [1][4].

$$P(e|w)=\sum_x P(x, e|w) \quad \text{(4.3)}$$

The variable x represents alignments between the individual chunks in the sentence pair and chunks in the sentence pair can be words or phrases.

Tuning is the process of finding the optimal weights for this linear model, where optimal weights are those which translation performance on a small set of parallel sentences [1][9]. Therefore, to find better weights, the machine translation system was tuned.

4.3.6. Decoding

Decoding is a method of finding the translation probability that maximizes the log-liner mode is exponential in the input sentence [4][6]. It starts by searching the phrase table for all possible translations and for all possible fragments of the given sources sentences. Decoder uses feature scores and weights to select the most likely translation [10]. It looks up all translations of every source word or phrase translation table and recombine the target language phrases that maximizes the translation model (after tuned) probability multiplied by the language model probability, which is,

$$w_{best} = \text{argmax}_e P(e|w) * P(w) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.6)$$

Taking English language as an input and displaying Wolaytta language as an output as illustrated in Figure 4.1.

4.4. Software's

The used Linux platform for this study, because of Moses developers recommended Linux platform (the Windows platform was not tested very well) [6]. Used different type of tools (software) for this study, such as Moses decoder, SRILM Language Modeling Toolkit, MGIZA++, and shell-scripts (normalization, tokenization, and lower case) for building the English to Wolaytta machine translation using statistical approach. All mentioned software are open source tools that are available on the Internet freely.

The other used software is Morfessor which helps to increase the performance of the system. It is a family of probabilistic machine learning methods for finding the morphological segmentation from raw text data [23]. In the morphological segmentation task, the goal is to segment words into morphemes which is the smallest meaning-carrying units. Morfessor is a family of methods for unsupervised morphological segmentation. It is used for the target language only.

Moses is a statistical machine translation system that allows us to automatically train translation models for any language pair [5]. It requires parallel corpus and provides different features, such as

- phrase-based which is the state-of-the-art in statistical machine translation that allows the translation of short text chunk and
- Factored which is words may have factored representation i.e. surface forms, lemma, part-of-speech, morphology, and word classes.

The Moses decoder developed by Marcu and Wong, they proposed for phrase-based [6]. It can be trained to translate between any languages pairs. Moses is the latest developments in the area of statistical machine translation research[1]. It is used in this study to produce English-Wolaytta translation.

One of the most difficult tasks in machine translation is to evaluate the output of the system. For this study, selected the BLEU as an evaluation metric. The BLEU metric is an IBM developed metric and well known for the machine evaluation for the machine translation [9] [10]. It checks how closer the candidate translation is to translation is to the reference translation based on the n-gram. BLEU score is based on the number of correct n-gram matches [20] [22].

CHAPTER FIVE

5. EXPERIMENT

5.1. Introduction

Under this chapter, the English-Wolaytta statistical machine translation system performance will be discussed in detail.

5.2. Experiment

As we mentioned in chapter four, the corpus are collected from different sources in different formats. All of the collected data was subsequently converted to plain text, clean up from the blank lines and noisy characters, and its encoding was converted to UTF-8 automatically to make it ready to train the system. In this study two groups of experiment conducted to come up with SMT for English-Wolaytta language pair. The first group of experiment focuses on unsegmented SMT, and the second group of experiment focuses on morphological segmented SMT, which are all developed using the freely available open source toolkits. These tools are: Moses used to train the system, SRILM for building the language model, MGIZA++ for word alignment. Morfessor 2.0 for morpheme segmentation and BLEU score for evaluation. Each group of experiment was conducted with six different corpora. Each corpora used 95% sentences for training, 2% sentences for tunning and 3% sentences for testing purpose. The following table summarizes the amount of bilingual data used in this research.

	English sentences	Wolaytta sentences
Corpus-1	5000	5000
Corpus-2	10000	10000
Corpus-3	15000	15000
Corpus-4	20000	20000
Corpus-5	25000	25000
Corpus-6	30000	30000

Table 5.1. Summary of the total sentences

5.2.1. Experiment-I: Unsegmented Corpus Set

This is the unsegmented group experiment conducted on English-Wolaytta language pair by using unsegmented parallel sentences. The experiment-I system have been trained using the

six different experiments. Out of 30,000 sentences parallel sentences, 28,500 sentences used for training, 600 sentences for tuning and the remaining 900 sentences for testing the system. The performance of the unsegmented experiment BLEU score is 8.46%.

	Sentences	Training set	Tunning set	Testing set
Corpus-1	5,000	4,750	100	150
Corpus-2	10,000	9,500	200	300
Corpus-3	15,000	14,250	300	450
Corpus-4	20,000	19,000	400	600
Corpus-5	25,000	23,750	500	750
Corpus-6	30,000	28,500	600	900

Table 5.2. Unsegmented experiment sets

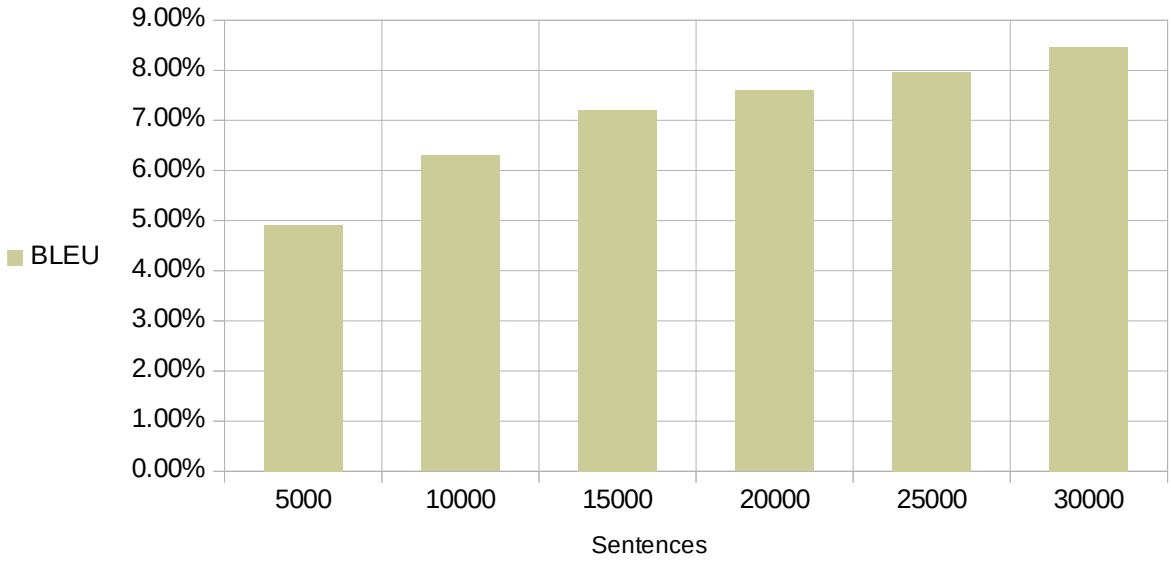

Figure 5.1. Unsegmented experiment BLEU scores

5.2.2. Experiment-II: Segmented Corpus Set

This is the second group experiment conducted on English-Wolaytta language pair by using morphological segmentation for the Wolaytta sentences only. The segmentation is performed using a segmentation model using Morfessor tool. The segmentation of the Wolaytta corpora has been before training, tunning and testing steps, using the created segmentation model as input for the unsupervised segmentation model.

Out of 30,000 sentences parallel sentences, 28,500 sentences used for training, 600 for tuning and the remaining 900 sentences for testing the system. The performance of the segmented experiment BLEU score is 13.21%.

	Sentences	Training set	Tunning set	Testing set
Corpus-1	5,000	4,750	100	150
Corpus-2	10,000	9,500	200	300
Corpus-3	15,000	14,250	300	450
Corpus-4	20,000	19,000	400	600
Corpus-5	25,000	23,750	500	750
Corpus-6	30,000	28,500	600	900

Table 5.3. Segmented experiment sets

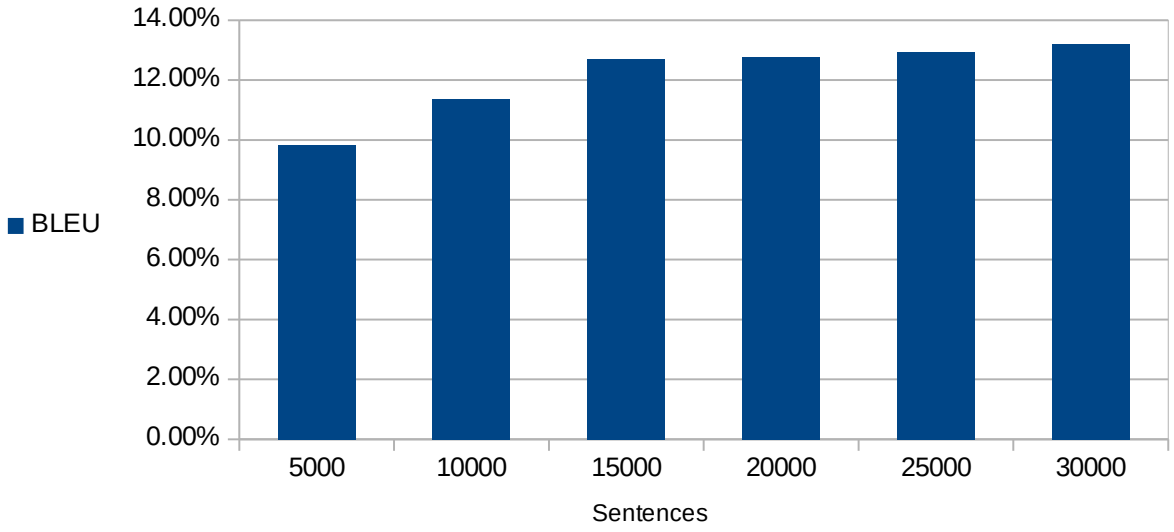

Figure 5.2. Segmented experiment BLEU scores

5.3. Discussion

The experiments are conducted by using two different experiments which are the unsegmented and segmented. Based on the first research question the experiments shown as the first English-Wolaytta SMT performance score 8.46% used unsegmented experiment and score 13.21% used segmented experiment.

Based on the second research question we have seen that better translation accuracy by increasing the size of the corpus for each experiment. These are 5000, 10,000, 15,000, 20,000, 25,000 and 30,000 corpus size (sentences) BLEU score 4.91%, 6.30%, 7.21%, 7.60%, 7.96% and 8.46% used unsegmented corpus and 9.83%, 11.38%, 12.70%, 12.77%, 12.93% and 13.21% used unsegmented corpus.

Based on the third research question the result of each experiments we have seen that the result recorded BLEU score shows that the segmented approach BLEU score 13.21% is better score than the unsegmented BLEU score 8.46% English-Wolaytta pair.

CHAPTER SIX

6. CONCLUSION AND RECOMMENDATION

6.1. Conclusion

The purpose of this study was developed of English-Wolaytta machine translation using statistical approach. In this research, experimentation of statistical machine translation of English to Wolaytta was conducted and a score 8.46% (unsegmented) and 13.21% (segmented) was found.

The development process of English-Wolaytta statistical machine translation involves collecting parallel corpus from the different sources and corpus preparation which also involves dividing the corpus for training set, tunning set and test set. The study began with a brief discussion of the Wolaytta language and how it differs from English language. It described the phrasal categories as well as the sentence structure of the Wolaytta and how the sentence structure affects the translation process with English because of the languages are using the different sentences structure. It also explained the articles, punctuation marks and conjunctions that are used in both languages.

Finally experiments were conducted by using the collected data set to test the accuracy and efficiency of the system by using two ways which are segmented and unsegmented experiments. The collected corpus divided by six corpus: Corpus-1 (5,000), Corpus-2 (10,000, Corpus-3 (15,000), Corpus-4 (20,000), Corpus-5 (25,000), and Corpus-6 (30,000) to answer the research questions. The first group experiment is conducted by using a unsegmented experiment and it has a BLEU score of 4.91%, 6.30%, 7.21%, 7.60%, 7.96% and 8.46%. The second group experiment is carried out by using morphological segmentation for the Wolaytta language to improve the performance of the system and it has a BLEU score of 9.83%, 11.38%, 12.70%, 12.77%, 12.93% and 13.21%. These experiments answer the second and third research questions by increasing the size of the corpus and morphological segmentation. Based on the experiments increasing the size of the corpus and morphological segmentation is a better way to improve the performance of the system.

Used a small amount of parallel sentences comparing from the very large corpus available languages like English, Germany, French, Chinese, Portuguese, and Indians[1]. The main reason for this is the scarcity of the Wolaytta language in digital format texts to train the system. Tried to improve the system by increasing the size of the bilingual corpus and using

morphological segmentation for the Wolaytta language. The obtained performance of the system is startup result for English-Wolaytta SMT.

6.2. Recommendation

The corpus taken for this study cannot be enough and a representative of the language, and future research should be conducted using a large set of corpus. Based on the above findings, the following areas could be explored further as a continuation of this study.

- Further researches in English to Wolaytta or Wolaytta to other languages machine translation should be performed and besides, a large corpus should be prepared in the Wolaytta language.
- Further results can be accomplished by increasing the size of the corpus set should be used to improve the performance of the system.
- All corpus used for this study is collected from Holly Bible, if the corpus prepared from different discipline or domain which represent the English-Wolaytta SMT in better way.

7. References

[1] Wikipedia, the free encyclopedia https://en.wikipedia.org

[2] L. Lessa, "Development Of Stemming Algorithm For Wolaytta Text", Msc thesis, Addis Ababa University, Ethiopia, 2003.

[3] M. Wakasa, " A Descriptive Study of the Modern Wolaytta Language", Doctoral Dissertation, Tokyo, Japan, 2008.

[4] W. J. Hutchins and H. L. Somers "An Introduction to Machine Translation". London: Academic Press, pp. ISBN: 0-12-362830-x, 1992.

[5] P. Koehn, Statistical Machine Translation, http//www.statmt.org, 2010

[6] S. Adugna, "English-Afaan Oromo Machine Translation:An Experiment using Statistical Approach", MSc thesis, Addis Ababa University, Ethiopia, 2009.

[7] Accredited Language Services, https://www.accreditedlanguage.com

[8] Statista – The portal for statistics, https://www.statista.com/

[9] E. Teshome, "Bidirectional English-Amharic Machine Translation: An Experiment using constrained corpus", MSc thesis, Addis Ababa University, Ethiopia, 2013.

[10] J. Daba, "Bidirectional English-Afaan Oromo Machine Translation Using:Hybrid Approach", Msc thesis, Addis Ababa University, Ethiopia, 2013.

[11] D. Beldados , "Automatic Thesaurus Construction From Wolaytta Text", Msc thesis, Addis Ababa University, Ethiopia, 2013.

[12] H. Fanta, "Speaker Dependent Speech Recognition For Wolaytta Language" ,Msc thesis, Addis Ababa University, Ethiopia, 2010.

[13] M. Yifiru, "Morphology-Based Language Modeling for Amharic", Ph.D Thesis, University of Hamburg, German, 2010.

[14] B. Adams, " A Tagmemic Analysis of the Wolaitta Language", Ph.D Thesis, University of London (unpublished), England, 1983.

[15] M. Wakasa, "A Sketch Grammar of Wolaytta", *Japan Association for Nilo-Ethiopian Studies*, 2014

[16] D. Dalke, " Tense, Aspect and Mood (TAM) in Wolayta", Msc thesis, Addis Ababa University, Ethiopia, 2012.

[17] Y. Solomon, "Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation", MSc thesis, Addis Ababa University, Ethiopia, 2017.

[18] Neural Machine Translation - Tutorial ACL 2016, https://sites.google.com/site/acl16nmt/

[19] K. Papineni, S. Roukos, T. Ward and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", July 2002.

[20] M. Gebreegziabher and L. Besacier. " English - Amharic Statistical Machine Translation", SLTU, 2012.

[21] T. Tsgaye, "English-Tigrigna Factored Statistical Machine Translation", Msc thesis, Addis Ababa University, Ethiopia, 2014.

[22] M. Hailegebreal, "Bidirectional Tigrigna-English Statistical Machine Translation", Msc thesis, Addis Ababa University, Ethiopia, 2017.

[23] S. Virpioja, P. Smit, S.A. Gronroos, and M. Kurimo, “Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline”, Aalto University publication series SCIENCE + TECHNOLOGY, Aalto University, Helsinki, ISBN 978-952-60-5501-5, 2013.