ENGLISH – AFAAN OROMOO MACHINE TRANSLATION: AN EXPERIMENT USING STATISTICAL APPROACH

A thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Information Science

BY

SISAY ADUGNA CHALA

ADDIS ABABA UNIVERSITY

APRIL, 2009

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

ENGLISH – AFAAN OROMOO MACHINE TRANSLATION: AN EXPERIMENT USING STATISTICAL APPROACH

BY

SISAY ADUGNA CHALA

Name and Signature of Members of the Examining Board

Ato Getachew Jemaneh, Chairman, Examining Board _____

Dr. Andreas Eisele, Advisor _____

Ato Ermias Abebe, Co - Advisor _____

Ato Sebsibe Hailemariam, Examiner _____

ACKNOWLEDGEMENT

First and foremost, I would like to pass my heartfelt thanks to my advisors Dr. Andreas Eisele and Ato Ermias Abebe without whose constant encouragement, advice and support this thesis would not have become a reality. I would especially thank Dr. Andreas Eisele and Prof. Hans Uszkoreit who allowed me to get access to computational resources in Saarland University (Germany) to conduct my research without which I would not have got sufficient computational infrastructure to conduct my research very well.

Secondly, I would like to thank Ato Mesfin Getachew, Dr. Rahel Bekele, Dr. Philipp Köehn, Dr. Sisay Fisseha, and Dr. Nega Alemayehu for their cooperative responses and hospitality by the time I was in need of their help.

Third, I would like to thank Ato Abebe Fite and Ato Dewano Kedir from Oromia Justice Bureau, Prof. Kevin Scannel, Mr. Zakia Posey, and Ato Amare Adugna for their support during my search for data.

Forth, I would like to thank the Deutscher Akademischer Austausch Dienst (DAAD) without whose financial support I would have stuck at a point of no return.

Last but not least, I would like to thank Aynalem T., Genet M., Mesfin W., Tsegaw K., Israel W. and Henock T. for their encouragement and constructive ideas and the good times we spent together. I would also thank those who have directly or indirectly contributed to the success of this thesis and whose names have not been listed here.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Acronyms

AAU – Addis Ababa University

ADJ – Adjective

BLEU – Bilingual Evaluation Understudy

CLIR – Cross-Language Information Retrieval

CSA – Central Statistical Authority

DAAD – Deutscher Akademischer Austausch Dienst

DT – Determiner

EM – Expectation Maximization (Estimation Maximization)

FDRE – Federal Democratic Republic of Ethiopia

MT – Machine Translation

NN – Noun

NP – Noun Phrase

SGML – Standard Generalized Markup Language

SMT – Statistical Machine Translation

SNNP – Southern Nations Nationalities and Peoples

SOV – Subject Object Verb

SRILM – Stanford Research Institute Language Modeling

SVO – Subject Verb Object

V – Verb

VP – Verb phrase

ABSTRACT

Machine Translation (MT) refers to the use of a machine for performing translation task which converts text in one Natural Language into another Natural Language. It can have many applications like cross-linguistic information retrieval and speech to speech translation systems. It can also assist professional translators by producing draft quality output that reduces cost that would be incurred if translation and typing was done manually from scratch.

English is the lingua franca of online information and Afaan Oromoo is one of the most resource scarce languages. For this reason, monolingual *Afaan Oromoo* speakers need to use documents written in other languages, among which English is the most popular one. To satisfy this need, translation of the English documents to Afaan Oromoo, and thus, making these online documents available in Afaan Oromoo is vital in addressing the language barrier thereby reducing the effect of digital divide.

Therefore, this thesis is focused on the development of a prototype English-Afaan Oromoo machine translation system using statistical approach, i.e, without explicit formulation of linguistic rules, as this approach involves low cost and swiftest way available these days. Using limited corpus of about 20, 000 bilingual sentences, a translation accuracy of 17.74% was achieved.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The goal of this chapter is to introduce the overall thesis. It highlights the background of the study, the statement of the problem and its significance, the objectives, methods, scope, limitations and applications of the research. The chapter concludes by showing the roadmap of the organization of the whole thesis.

1.2 Background

Communication is a vital part of personal life and is also important in business, education, and any other situation where people encounter each other (Microsoft Encarta Encyclopedia, 2004). Language (be it verbal, non-verbal, or written) is the principal means of communication. *Afaan Oromoo* (also referred to as Oromiffa, Oromigna or Oromo) is one of the many languages spoken by the speakers of the language for the purpose of communication.

According to Microsoft Encarta Encyclopedia (2004), *Afaan Oromoo* is a mother tongue of more than 17 million people in Ethiopia. The language is also spoken in other east African countries like Kenya and Somalia (Tilahun, 1993). It is also a medium of instruction and a school subject in primary and secondary schools in the Oromiya regional state, one of the administrative regions that has the largest population of all the regions of the Federal Democratic Republic of Ethiopia (FDRE). According to the statistical abstract of the Central Statistical Authority (CSA), in 2002, the total population of Oromiya region is 24,395,000 followed by Southern Nations/Nationalities and Peoples (SNNP) region and Amhara region with a

population of 17,669,000 and 13,686,000 respectively. This statistical data shows that *Afaan Oromoo* is the official working language of the Oromiya region that has a population of more than 24 million. All of these reasons make the language to be predominantly used in different offices. Since *Afaan Oromoo* writing in Latin script began only in 1991 (Tilahun, 1993), there is insufficient document prepared in this language.

On the other hand, English is the lingua franca of online information (Hersh, 2003). Hersh (2003) added that most international scientific conferences and publications use English as their required language. An analysis of the web characterization project of the Online Computer Library Center (www.oclc.org) has found that 71% of all web pages are in English. The next most common languages are German (7%), Japanese (6%), French (3%), and Spanish (3%). The Internet contains useful documents like news items in English which are inaccessible for the *Afaan Oromoo* speakers due to lack of English language. Therefore, translation of documents from English to local languages such as *Afaan Oromoo* is necessary for utilizing these useful online documents for local use. So, what is translation?

Smith (2008) defined translation as an act of interpretation of the meaning of a content and consequent re-production of equivalent content. The content or the text that is required to be translated is called "Source Text" (English in this case) and the language into which the source text is to be translated is known as "Target Text" (*Afaan Oromoo* in this case). There are two possibilities to translation, namely: Manual Translation (in which any translation task is carried out by human translators) and Automatic or Machine Translation (in which any translation task is carried out by computer software).

Thus, the focus of this research is on automatic or machine translation from English to Afaan Oromoo.

1.3 Statement of the Problem and Justification

According to the Web Characterization Project of the Online Computer Library Center (www.oclc.org), there are abundant documents in English on the Internet. However, lack of English language knowledge creates a problem of fully utilizing these documents. The researcher believes that studying how to make these documents available in local languages (such as *Afaan Oromoo*) is vital in addressing the language barrier thereby reducing the effect of digital divide.

There is scarcity of documents written in *Afaan Oromoo* since its writing in Latin script began only in 1991 (Tilahun, 1993). For this reason, *Afaan Oromoo* speakers, libraries, schools, offices and court houses need to use documents written in other languages. Among these languages English is the most popular one. To satisfy this need, translation of the English documents to Afaan Oromoo is important. This can be achieved through manual or automatic translation as explained in the background section.

Manual translation is a slow and expensive process and also requires the translator to be a professional in specialized field (Smith, 2008). It has been regarded by many translators as a repetitive, monotonous and thus boring, but at the same time, difficult job (Bernard, 1998).

Although machine translation (MT) will not achieve the pinnacles of human translator's art (at least at the current level of technology), it would minimize cost and maximize speed, both of which are the current demands of the business community.

It has repeatedly been proven that machine translation aids human translators to do their job more efficiently. Locke (1955) puts this statistically that machine translation increases the productivity of human translators by 30% or more.

MT also allows documents to be translated that would otherwise remain untranslated due to lack of resources. For example, Wiki in Afaan Oromoo (http://om.wikipedia.org) is almost a collection of empty web pages while there are equivalent articles in English that would fill those empty pages and would be useful. If MT platform for wiki were in place, those articles available in English would be available in Afaan Oromoo too.

For the automatic translation to be a reality, studying the language in question (Afaan Oromoo) is crucial. From the review of the number of works done on the linguistics aspect of the Afaan Oromoo language, it seems that the language is being studied intensively. However, corresponding development in the computational aspect is very rare, though there are few research works (Wakshum, 2000; Morka, 2001; Diriba, 2002; Asefa, 2005) initiated to address the problem. Oromosoft (www.oromosoft.com), a company based outside Ethiopia is also working on office software tools for *Afaan Oromoo*.

Apart from these, to the best of the researcher's knowledge, no research has been and is being conducted directly on English-*Afaan Oromoo* Machine Translation to the date of this thesis preparation.

Thus, it is worthy to research on MT to make machine translation system available for *Afaan Oromoo* speakers (to let them be able to use the documents prepared in

English without renouncing their own language), thereby reducing the impact of digital divide that would result due to language barrier.

1.4 Objectives of the Study

1.4.1 General Objective

The general objective of this study is to develop a prototype English-*Afaan Oromoo* machine translation system using statistical approach, i.e, without explicit formulation of linguistic rules.

1.4.2 Specific Objectives

With the aim of achieving the general objective specified above, the specific objectives that would be carried out in the study are:

- to identify the syntactic relationship between English and *Afaan Oromoo* languages.
- to discuss the different techniques and approaches employed so far in machine translation tasks and select the one that seem appropriate to this specific case.
- to prepare and organize training and test data.
- to adopt/develop a machine translation system that employs statistical machine learning algorithm.
- to train the machine translation system adopted/developed with training corpus data.
- to test the effectiveness of the machine translation technique developed.
- to discuss and report the experimental results found.

1.5 Research Methodology

In order to achieve the objectives of this research, different methodologies were employed. The following subsections discuss the methodologies that were followed in this study.

1.5.1 Literature Review

Developing English to *Afaan Oromoo* machine translation system requires review of syntactic structure of both languages. In addition to studying the syntactic structure of the individual languages, the syntactic relationship (similarities and differences) between them also needs to be studied. The typological facts about cross-linguistic similarities and differences that were studied include word order of noun, verb and objects in simple declarative clauses (Jurafsky and Martin, 2008). For example, in English, a simple declarative sentence is in Subject-Verb-Object (SVO) order while in Afaan Oromoo it is in Subject-Object-Verb (SOV) order. Yet another typological fact is the word order of noun and adjective. For example, in English, nouns follow adjectives (as in *tall man*) while in Afaan Oromoo the reverse is true (as in *namicha dheeraa*). Here *dheeraa* is an adjective and it means 'tall' and *namicha* is a noun and it means 'man'. The researcher believes that these cases have something to do in the tasks of word alignment, language modeling and decoding.

Since there are different approaches used in machine translation, literature review was also done on the approaches. The approaches and the different algorithms used in implementing them were studied thoroughly. In-depth exploration of the recent developments in MT in general and SMT in particular was made.

In the literature review, the previously conducted researches that are related to this research were also explored.

1.5.2 Data Collection

The researcher tried to get data from news agencies and failed to get any parallel data. Therefore, the researcher made use of existing and publicly available translated documents from other sources. These documents include the Constitution of FDRE (Federal Democratic Republic of Ethiopia), Universal Declaration of Human Right, proclamations of the Council of Oromia Regional State, religious documents, and other documents as these are the already translated and available documents. The collected documents were then divided into training set and testing set in such a way that nine-tenth of it was used for training and the rest (one-tenth) was used for testing the system. The choice of the proportion is arbitrary though it is intentionally done to make the training corpus larger size.

1.5.3 Data Preprocessing

Documents were preprocessed so as to fit the format the modeling tools require. This includes breaking of the document into sentences in such a way that separate sentences be on separate lines and corresponding parallel documents being on different files with corresponding sentences on corresponding lines. For this purpose, scripts that perform splitting documents into sentences were written using PERL (Practical Extraction and Report Language). The researcher chose PERL because it is the most efficient language (in terms of programmer's time) he ever knows for string processing, for working with regular expressions, and that most of the scripts like the one used for sentence alignment are written in this language. In addition, scripts for format conversion and unifying encoding were written in python.

1.5.4 MT Approach

Machine Translation can be done in different approaches. Out of these, rule-based and statistical approaches are the major ones. Rule-based approach, as its name shows, is based on linguistic rules ("what should be") whereas statistical approach pioneered by IBM (Brown et al., 1990) is based on data ("what really is"). In this research, the later approach (statistical machine translation or SMT) was used. The reason why statistical approach was chosen is that this approach does not require linguistic preprocessing of the documents, except that it requires translation of the text. SMT is basically data-driven, i.e., it learns from the data distribution itself. All it needs is two things basically. The first is parallel corpus (i.e., document in source language and its equivalent translation in the target language) for translation modeling. The second is monolingual corpus (i.e., document in target language) for language modeling.

For word-alignment, Expectation Maximization (EM) Algorithm (Koehn, 2006), one of the statistical learning algorithms that employ Baye's Rule, was used to conduct the experiment of the translation. Word alignment was obtained using GIZA++ (Och and Ney, 2003) as it is the implementation of EM and is open source tool. The translation experiment was conducted using the adopted tool.

In this research, the researcher used n-gram (Manning and Schütze, 1999) language model because Crego et. al. (2005) reported that n-gram approach outperforms the phrase-based one. The n-gram language model was trained with the SRILM language modeling package (Stolcke, 2002). The reason for choosing this toolkit was its public availability.

Decoding was done with Moses (Koehn et al, 2007) and the software included with it. Here also the reason for using this tool was its public availability.

Performance of the system was measured by the quality of its output. This was done automatically using BLEU (Bilingual Evaluation Understudy) metric (Papineni et al, 2002). Finally the result was discussed and reported.

1.6 Application of the Result

The result of this research is machine translation system prototype and it is used to get English documents translated into *Afaan Oromoo*. It can have many applications. First, it will reduce the language barrier that would hinder the monolingual *Afaan Oromoo* speakers from utilizing documents prepared in English and it will reduce the problem that would arise from scarcity of documents in offices, libraries, schools and court houses.

Second, machine translation is used in CLIR (Croft and Lafferty, 2003). In CLIR, machine translation is applied to translate the query or the result of the CLIR system.

Third, Schultz and Kirchhoff (2006) explained that machine translation is one of the requirements for speech to speech translation systems. For automatic speech to speech translation, it is used to translate the output of speech-to-text subcomponent to produce text that will be used as input for text-to-speech subcomponent.

Finally, machine translation is used to assist professional translators by producing draft quality output that reduces cost that would be incurred if translation and typing was done from scratch.

Therefore, beneficiaries of the result of this research includes: individuals, libraries, businessmen (translators), schools, offices, court houses, and researchers.

1.7 Scope and Limitation of the Study

1.7.1 Scope of the Study

The scope of the study is limited to building a prototype English-Afaan Oromoo MT system as it is not possible to build a full-fledged system with the time and budget allotted for this research. The prototype English – Afaan Oromoo MT system is the customization of the Moses (Koehn et. al., 2007) open source system that was built for Indo-European languages and is limited to translating English text to Afaan Oromoo text.

1.7.2 Limitation of the Study

There were different limitations that were faced during the process of conducting this research. The major ones are lack of educational material, data, computational infrastructure, and finance. The lack of relevant educational materials like appropriate books and journals that are required by a multidisciplinary research like this one posed an insurmountable limitation.

A limitation pertaining to data is the absence of sufficient amount of machine readable documents in Afaan Oromoo that made the research not capable of producing high quality output. Although there is some amount of monolingual corpus in Afaan Oromoo, it does not have its English equivalent. As a result of this, the research is limited to working with very little bilingual text.

The computational resources and finance have been extended beyond what was allotted by Addis Ababa University for this particular research by using the

computational infrastructure abroad, and by gaining extra financial support from DAAD respectively.

1.8 Organization of the Thesis

The thesis is organized into six chapters. This chapter discusses the background, statement of the problem and its significance, objective, scope and limitation, methodology, and the application of the result of the study.

The second chapter presents the lexical, syntactic and semantic similarities and differences between the source language (English) and the target language (Afaan Oromoo). In other words, chapter two briefly explains some of the language aspects of Afaan Oromoo and their corresponding ones in English that influence the process of translation.

The third chapter discusses the field of machine translation. In chapter three, brief explanation of the field of MT, its development, and the different approaches employed since the beginning of the discipline to date is made.

The fourth chapter deals with the statistical machine translation as it is the approach that will be used in this research. The chapter discusses the different methods that involve statistical modeling for machine translation.

In chapter five the architecture and the general overview of the system is described. The design of the system together with the components of the software tools that are used and their relationship are also explained in detail. The training, testing and analysis of the result are also reported in the same chapter. The method of experimentation, the parameters of the experimentation, the results of the experimentation, and the analysis of the result of the experimentation are also discussed in chapter five.

In chapter six, the current work is concluded, i.e., what has been found and learned from this research is summarized. Besides, the future works that are identified and believed to augment this system are also recommended.

CHAPTER TWO

CHALLENGES IN ENGLISH-AFAAN OROMOO MT

2.1. Introduction

In translating documents from one language to another, ambiguities of different forms are inevitable. In order to understand some of the ambiguities and appreciate what linguistic information can be extracted automatically from the data without explicit formulation of rules, it is important to see the lexical, syntactic, and semantic similarities and differences between the two languages in question. This chapter is dedicated to reviewing the linguistic relationships between English and Afaan Oromoo.

2.2. Afaan Oromoo Alphabet

Since 1991 (Tilahun, 1993), Afaan Oromoo writing uses the alphabetic writing system for many reasons, among which is adaptability to computer. Afaan Oromoo writing is straightforward, i.e., it is written as it is read and read as it is written.

In Afaan Oromoo alphabet, there are ten vowels (five of them are short vowels and the rest are long vowels) and twenty-four basic consonants (18 of them are single-character and the rest are diphthongs-letters having two characters like 'ch') as shown in Table 2.1, Table 2.2 and Table 2.3 below.

Short vowels	Pronounced as	Long Vowels	Pronounced as
A	u in but	aa	a in bag
E	a in lake	ee	a in late
I	i in big	ii	ee in week
O	o in got	oo	o in doll
U	u in shoot	uu	oo in book

Table 2.1 Vowels in Afaan Oromoo (Tilahun, 1989)

Letter	Pronounced as	Letter	Pronounced as
b	ba in bad	m	ma in man
c[3]		n	na in narrow
d	da in dad	q[4]	
f	fa in fat	r	ra in rabbit
g	Ga in God	s	sa in sad
h	ha in hard	t	ta in target
j	ju in jump	w	wa in
k	ca in cat	x[5]	
l	la in large	y	ya in yard

Table 2.2 Basic Consonants in Afaan Oromoo (Tilahun, 1989)

Letter	Pronounced as
Ch	ch in charge
Dh[1]	
Dz	s in vision
Ny	ñ in Español
Ph[2]	
Sh	sh in shark

Table 2.3 Diphthongs (Tilahun, 1989)

Since there are not any indigenous Afaan Oromoo words that contain 'p', 'v', and 'z', the basic alphabet also does not include these. However, when foreign words such as 'police' are refereed to, these characters are used. Thus, the extended alphabet of the Afaan Oromoo includes these characters ('p', 'v', and 'z') to support foreign words.

In addition, apostrophe is used to represent a sound in addition to its use as punctuation mark. It represents a hiccup-like sound (called *hudhaa*) as in *re'ee* and *fal'aana*.

2.3. English - Afaan Oromoo Linguistic Relationship

English and Afaan Oromoo have some structural and typological differences as well as similarities. The syntactic and lexical relationships which the researcher believes to be interesting in English –Affan Oromoo translation are discussed below.

[1] No Equivalent sound in English. It is a very difficult sound to produce for non-native speakers. It is like saying 'd' and 'a' at the same time. Non-native speakers of the language use 'd', instead.
[2] No Equivalent sound in English but it is closer to 'p'
[3] No equivalent sound in English but it is closer to 'ch'
[4] No equivalent sound in English but it is closer to 'k'
[5] No equivalent sound in English but it is closer to 't'

2.3.1. Word Order

English and Afaan Oromoo have differences in their syntactic structure. Particularly, the syntax of English is SVO whereas that of Afaan Oromoo is SOV in a simple declarative sentence. Moreover, in English, adjective precede the noun they modify while the reverse holds for Afaan Oromoo, nouns precede their adjectives. For example, in *mucaa dheeraa* (tall boy), *dheeraa* (adj) precedes *mucaa* (noun).

2.3.2. Noun

Nouns in Afaan Oromoo can vary to reflect number, gender, and case (subjective, objective or possessive).

Number of noun can be indicated by using numerals and quantifiers. Usually, number in countable nouns is indicated by numerals and quantifiers, for example, in *sangaa shan* (five oxen), *shan* (five) indicates the number of oxen and in *sangoota baay'ee* (many oxen) *baay'ee* (many) is the quantifier to reflect plurality.

In addition to numerals and quantifiers, plural markers also indicate the number of noun. Unlike English that mainly has a single plural marker (suffixing 's' to a noun), Afaan Oromoo has many plural markers (Tilahun, 1989) such as *-oota*, *-olii*, *-lee*,and -w*wan*. In Afaan Oromoo, a plural noun is produced by suffixing these plural markers on the singular noun. For example, suffixing *–oota* to a singular noun, a plural can be formed as in, *nama* (person) –> *namoota* (persons). Similarly, suffixing *–olii* to some singular nouns changes the singular to plural as in, *raammoo* (worm) –> *raammolii* (worms).

Unlike languages which assume gender for every noun (for example, French), Afaan Oromoo considers gender only for nouns which are naturally gender oriented (genetically male or female), i.e., only animals. Gender of nouns can be indicated in different ways: by using differentiated lexical items as in *abbaa* (father) and *haadha*

(mother), or genetic terms denoting usually animals mark gender by adding *kormaa* for masculine and *dhaltuu* for feminine as in *saree kormaa* (male dog) and *saree dhaltuu* (female dog) or by using morphological difference as in *jaarsa* (old man) and *jaartii* (old woman), *obboleessa* (brother) and *obboleetti* (sister).

With respect to case, nouns have three major cases: nominative (subjective), objective, and possessive. The common nominative case markers in Afaan Oromoo are: *–n*, *-ni*, and *–i*. These markers inflect the noun and finally form the subjective case as in,

*mucaa**n** du'e* (the child died)

*nam**ni** dhufe* (a person came)

*marg**i** bikile* (the grass grew)

In Afaan Oromoo, noun in the direct object is unmarked as in,

*Inni **mucaa*** waame (He called the child)

*Inni **nama*** jaalata (He loves people)

whereas noun in the indirect object is marked by inflecting it with *–f* or *dhaaf*, and *-tti* or *-tiif* as in,

*Kitaaba kana mucaa**f** (mucaa dhaaf) kenni.* (Give this book to the child)

*Kitaaba kana na**tti** kenni.* (Give this book to me)

*Kitaaba kana mucaa kee**tiif** kenni.* (Give this book to your child)

There are two major ways of forming possessive case in Afaan Oromoo. First, possessive can be formed by juxtaposing two nouns in genitive relationship (i.e., putting the possessor after the possessed noun) as in,

Mana Guutaa (Guta's house)

Second, possessive can be formed by prefixing kan to nouns as in,

Kan Guutaa (Guta's)

2.3.3. Pronouns

Afaan Oromoo is different from English in the way it treats the second person pronoun. In English, there is one representation of the second person pronoun (you) to refer to the plural and singular nouns, or to refer to formal in the subject form, possessive form as well as in the object form. However, in Afaan Oromoo, the second person plural and the formal second person are the same (*isin*) whereas the singular subject form is (*ati*) and the singular object form is (*si*) as shown in the following examples,

Subjective form:

You (singular, informal) chased the blue cat = *Ati adurree cuquliisa reebde.*

You (plural) chased the blue cat = *Isin adurree cuquliisa reebdan.*

You (formal) chased the blue cat = *Isin adurree cuquliisa reebdan.*

Objective form:

The blue cat chased you (singular, informal) = Adurreen cuquliisni si reebde.

The blue cat chased you (plural or formal) = Adurreen cuquliisni isin reebde.

Possessive form:

The blue cat is yours (singular, informal) = Adurreen cuquliisni keeti.

The blue cat is yours (plural or formal) = *Adurreen cuquliisni keessan.*

Here, the English word 'you' can be translated to Afaan Oromoo words '*ati*', '*isin*' or '*si*' depending on the context of use. Likewise, the English word 'yours' can be translated to the Afaan Oromoo words '*keeti*' or '*keessan*'. This lexical divergence affects the performance and result of the translation system.

Another major issue related to pronoun in English – Afaan Oromoo MT is the lexical divergence that occurs in demonstrative pronouns. For example, the demonstrative pronoun 'this' has different forms when it refers to subject, direct object and indirect object for both genders as shown in the table below.

Form	This		That	These	Those
	Masculine	Feminine			
Subject	*kun*	*tun*	*Sun*	*kunnin*	*sunniin*
Direct Object	*kana*	*tana*	*San*	*kanniin*	*sanniin*
Indirect Object	*kanaa*	*tanaa*	*Sanaa*	*kenniinii*	*sanniinii*

Table 2.4 The lexical divergence of demonstrative pronouns

Therefore, when translating the demonstrative pronoun 'this' from English to Afaan Oromoo, there is a difficulty of choosing one of '*kun*', '*kana*', '*kanaa*', '*tun*', '*tana*', or '*tanaa*' likewise for 'that', 'these', and 'those'.

As a result of the above personal or demonstrative pronouns, choosing the correct target word is a difficult problem for English – Afaan Oromoo machine translation because a single word has multiple possible translations. The selection of the most appropriate translation is context-dependent.

2.3.4. Adjectives

Adjectives in Afaan Oromoo reflect the gender of the noun they modify, although there are some adjectives which take the same form in both genders. Here are some examples:

jaartii hiyyeettii (poor old woman), *jaarsa hiyyeessa* (poor old man),

sa'a adii (white cow), *qotiyyoo adii* (white ox)

In addition, Afaan Oromoo adjectives reflect the number (plurality or singularity) of the noun they modify. Adjectives usually form their plurals in two ways. Firstly, they form their plurals by duplicating the first syllable and geminating the initial consonant in the first syllable, for example,

nama dheeraa (tall man) –> *namoota dhedheeraa* (tall men).

Secondly, they form their plurals by using the suffix *–oota*, for example,

dhukkubsataa (sick) –> *dhukkubsatoota* (the sick)

2.3.5. Prepositions and Postpositions

Both postpositions and prepositions exist in Afaan Oromoo. Prepositions which normally precede nouns or pronouns in English are in most cases placed after the noun or pronoun they modify in Afaan Oromoo, i.e., prepositions in English become postpositions in Afaan Oromoo. For example, *On* earth means lafa *irra* where the English preposition *on* is translated to the Afaan Oromoo postposition *irra*.

Nevertheless, some prepositions in English remain prepositions in Afaan Oromoo. For example,

I went *to* Arsi. (ani *gara* Arsii deemee ture.)

In addition, some of the prepositions in English may be indicated by morphology or can stand by themselves, for example, the English preposition 'on' can stand by itself as '*irra*' or can be suffixed to a noun as '*-rra*' as in,

Lafa irra (lafarra) = On earth

Similarly, 'from' can stand by itself as '*irraa*' or can be suffixed to a noun as '*-rraa*' as in,

Lafa irraa (lafarraa) = from earth

2.3.6. Conjunctions

Among the three common conjunctions (and = *fi*, or = *yookin*, but = *haa ta'u malee*), only *but* creates a problem as it generated three different words in Afaan Oromoo. The following examples use the three common conjunctions:

Right and left = *Mirgaa fi bitaa*

Right or left = *Mirga yookin bitaa*

She is thin but strong = *Ishiin qall'oo dha haa ta'u malee jabduu dha*.

The Afaan Oromoo conjunction "fi" appears in documents in two forms: it may stand by itself or it may be attached with the first word that involves in the conjunction as in *kanaa fi san* or *kenaaf san* both of which mean *this and that*.

2.3.7. Verbs

Auxiliary and modal Verbs in Afaan Oromoo follow lexical (standard) verbs, but in English the reverse is true. The infinitive (two different words in English) is formed by suffixing *–uu* to the stem (root) of the verb. For example, *dhufuu* (to come), *deemuu* (to go).

Afaan Oromoo active and passive voices are also slightly different from that of English. In active voice, the root verb is inflected to show the number and gender of the doer of the action as in,

He cut wood. = *Inni muka mure.*

They cut wood. = *Isaan muka muran.*

You cut wood. = *Isin muka murtan.*

Passive voice is formed by suffixing *–ame, -amte, -aman, -amtan, -amne* to the root verb depending on the number and gender of the receiver of the action as in,

Wood is cut. = *Mukni murame*

Woods are cut. = *Mukoonni muraman*

You (plural or formal) are beaten. = *Isin tumamtan*

We are beaten. = *Nuti tumamne*

I am beaten. = *Ani tumame*

They are beaten. = *Isaan tumaman*

2.3.8. Adverbs

In English, adverbs follow verbs they modify. However, Afaan Oromoo adverbs precede the verb they modify as shown in the following example.

Ishiin suuta deemti. (She walks slowly)

Inni boru dhufa. (He will come tomorrow)

2.3.9. Articles

Unlike Afaan Oromoo, English often requires articles to appear before nouns. In English, there are three main semantic choices for article insertion: no article, indefinite article (a, an, some, any) and definite article (the). The fact that Afaan Oromoo has no articles before nouns makes the translation of noun phrases difficult because in English some nouns take articles others do not. However, Afaan Oromoo varies nouns morphologically in such a way that the last vowel of the noun is dropped and suffixes (*–icha, -ittii, -attii*) added to show definiteness instead of using definite articles. For example,

the cat (masculine) = *adurree + icha (adurricha)*

the cat (feminine) = *adurree + attii (adurrattii)*

furdoo (feminine) = *furdoo + ittii (furdittii)*

2.3.10. Punctuation Marks

All punctuation marks that are used in English are also used for the same purpose in Afaan Oromoo except the apostrophe. Unlike its use to show possession in English, it is used as a symbol to represent a hiccup (called *hudhaa*) sound in Afaan Oromoo writing. This influences the tokenization step of the preprocessing phase that leads to breaking the word into two separate (most probably meaningless) pieces using the apostrophe as a boundary. As a consequence, the MT system performs poorly if due care is not given to this punctuation mark.

CHAPTER THREE

MACHINE TRANSLATION

3.1. Introduction

The purpose of this chapter is to review the field of machine translation. It covers what machine translation is, historical development of machine translation, the different approaches employed in machine translation and the challenges in machine translation.

3.2. Machine Translation

Machine Translation (MT) is an application of Artificial Intelligence that is concerned with extracting the meaning of one language (source language) and consequent generation of another language (target language). In its full generality, MT is the use of computers to automate part or all of the process of translating document from one language to another.

MT can be used in various scenarios in real life. Jurafsky and Martin (2008) summarized the applications of MT in the following two general categories: applications of MT by tasks (purposes), and applications of MT by the number and direction of translation. The following subsections discuss these categories.

Applications of MT based on tasks (purposes)

Based on the task and purpose of the translation, applications of MT can be classified into the following four categories: rough translation, restricted source translation, pre-edited translation, and literary translation.

In the case of *rough translation*, MT is applied in two situations (Jurafsky and Martin, 2008; Russell and Norvig, 2003). First, when only the rough translation is adequate, i.e., in a situation where the goal is just to get the gist of a passage. In such cases,

ungrammatical and inelegant sentences are tolerated as long as the meaning is clear. For example, in web surfing, a user is often satisfied by a rough translation of foreign web page. Second, when human post-editor is used, it is also called Computer Assisted Human Translation (CAHT) or Computer Assisted Translation (CAT). This is good for high volume translation jobs and those requiring quick turn-around time like localization of software or hardware manuals. This type of system saves money because the rough output can be post-edited by a monolingual human without having read the source and such editors can be paid less as compared to bilingual translators.

Restricted source translation (Russell and Norvig, 2003) also known as small sublanguage domain (Jurafsky and Martin, 2008) is a case in which the subject matter and format of the source text are severely limited. In such cases, fully automatic high quality translation is achievable (Jurafsky and Martin, 2008). One of the most successful examples is the TAUM – METEO (Chevalier et. al., 1978) system, which translates weather reports from English to French. It works because the language used in weather report is highly stylized and has some well-known structure, i.e., it has very low ambiguity.

Pre-edited translation is a translation in which a human pre-edits the source document to make it conform to a restricted subset of the source language before MT. This approach is particularly cost-effective for one-to-many translations as is the case for companies that sell the same product in many countries to prepare product documentation, user manuals, etc.

Literary translation is the case in which all the nuances of the source text are preserved. This is currently beyond the state of the art for MT (Russell and Norvig, 2003) as it requires subjective judgment of cultural relationships.

Applications of MT by number and direction of translation

Application of MT can also be characterized by the number and direction of the translation as: one-to-many, many-to-one, and many-to-many translation (Jurafsky and Martin, 2008).

Translation tasks that require *one-to-many* translation include localization of computer manuals, and web pages from English to other languages all over the world as English dominates the language of the web and technology. This research focuses on the translation from English to Afaan Oromoo which can be the subset of one-to-many translation.

Many-to-one is relevant for anglophone readers who need to get the gist of web content written in other languages.

Many-to-many is required in environments like the European Union, where eleven official languages need to be intertranslated.

3.3. History of MT in Brief

The beginning of MT may be dated to a memorandum written in March 1947 from Warren Weaver of the Rockefeller Foundation to cyberneticist Norbert Wiener (Hutchins, 1986) which contains the following two sentences as has been taken from Arnold et. al. (1994).

> *"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."* Warren Weaver

From the mid-1970s onwards, the demand for MT came from quite different sources with different needs and different languages. The demand was towards cost-effective

machine-aided translation systems that could deal with commercial and technical documentation in the principal languages of international commerce.

The 1980s witnessed the emergence of a wide variety of MT system types, and from a widening number of countries. Examples include, Systran (Toma, 1977) operating in many pairs of languages, Logos (Bernard, 1998) (German-English and English-French), the internally developed systems at the Pan American Health Organization (Leon, 1984) (Spanish-English), the Metal (Slocum, 1984) system (German-English) and major systems for Japanese-English translation from Japanese computer companies.

The end of the 1980's was a major turning point for MT for two reasons. Firstly, a group from IBM published the results of experiments on a system called Candide (Brown et. al., 1990) based purely on statistical methods. Secondly, certain Japanese groups began to use methods based on corpora of translation examples, i.e. using the approach now called 'example-based' translation. In both approaches the distinctive feature was that no syntactic or semantic rules are used in the analysis of texts or in the selection of lexical equivalents; both approaches differed from earlier 'rule-based' methods in the exploitation of large text corpora instead of using a set of hard-coded rules.

In the 1990's, the use of MT and translation aids by large corporations has grown rapidly – a particularly impressive increase is seen in the area of software localization (i.e. the adaptation and translation of equipment and documentation for new markets). On the research front, the principal areas of growth are seen in example-based and statistical machine translation approaches, in the development of speech translation for specific domains, and in the integration of translation with other language technologies.

In 2000's, the move is towards combining the rule-based and SMT paradigms (Eisele

et. al., 2008; Nogués, 2006) with the goal of improving the quality of the output as well as the performance of the system.

3.4. Approaches to Machine Translation

Basically, there are two different methods for carrying out machine translation based on the requirements to do translation, namely: Rule-based approach (discussed in section 3.4.1) and Corpus-based approach (discussed in section 3.4.5).

3.4.1. Rule-based Approaches

Rule-based machine translation systems are fundamentally based on formulated rules for translation. According to Jurafsky and Martin (2008), there are three classical systems categorized by how they perform translation, namely: direct approach, transfer approach, and Interlingua approach. The following sections briefly discuss the different approaches used in Rule-based MT.

3.4.2. Direct Approach

Direct translation is performed by considering individual words in the source language text, translating each word as it goes. 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, simple reordering rules can be applied, for example, for moving adjectives after nouns when translating from English to Afaan Oromoo.

Figure 3.1 depicts the processes employed in the direct approach. From the diagram, the largest oval shows that the major component in the direct approach is the bilingual dictionary.

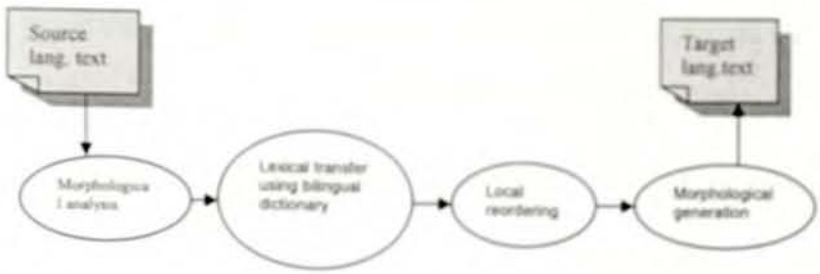


Figure 3.1 Direct Translation Systems

Using direct approach, the English sentence,

"The blue cat chased the brown mouse."

is equivalent (in word-for-word translation) to the following Afaan Oromoo sentence.

"Cuqulisni adurre reebde magaala hantuuta."

Now, reordering of adjectives and nouns yields:

"Aduree cuqulisni reebde hantuuta magaala."

As it can be seen from the above example, direct approach is too focused on individual words, that is, though it can handle single-word reordering, it cannot handle longer-distance reordering (such as changing from SVO to SOV) or those involving phrases or larger structures (Jurafsky and Martin, 2008).

3.4.3. Transfer approach

In transfer approach, the input text is first parsed and then rules to transform the source language parse structure into the target language structure are applied. Then the target language sentence is generated from the parse structure. Transfer approach overcomes the language differences by adding structural and phrasal knowledge to the limitations of the direct approach. Generally, English has SVO while Afaan Oromoo 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. For example, let S represent the English sentence:

"*The blue cat chased the brown mouse*." First, the SVO is changed to SOV as shown in the figure below.

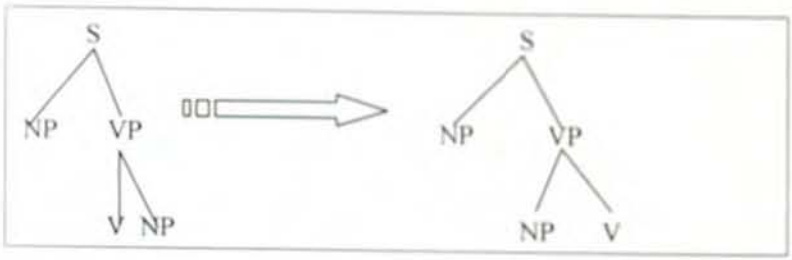


Figure 3.2. A simple syntactic transfer (Subject-Verb-Object)

Thus, the sentence becomes:"*The blue cat the brown mouse chased.*" Then the adjectives and nouns are interchanged.

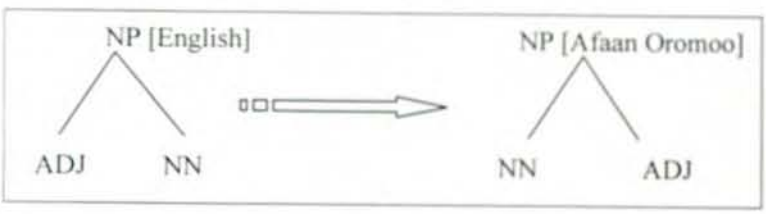


Figure 3.3. A simple syntactic transfer (adjective and noun)

This yields, *"The cat blue the mouse brown chased."* Now, lexical transfer can take place to yield: *"Adurreen cuquliisni hantuuta magaala reebde."* The following figure shows the tree transformation when translating the given sentence.

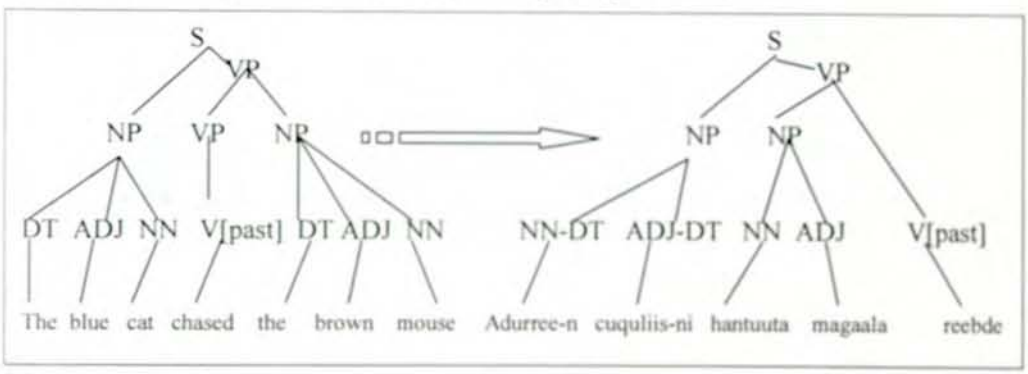


Figure 3.4. English to Afaan Oromoo Sentence translation using Transfer approach

3.4.4. Interlingua Approach

In the Interlingua approach, the source language text is analyzed into an abstract meaning representation called an Interlingua. Then the target language text is generated from this representation. This approach is appropriate for many-to-many translations as it reduces the language pair interdependence, i.e., each language is dependent on the meaning (Interlingua) and not on another language. As can be seen from Figure 3.5 below, addition of one more language (say German) into the system requires only the German Analysis & Generation module. In the Interlingua approach, there are n bidirectional systems and one representation of the meaning for n languages to be intertranslated. This makes the Interlingua approach far better than the previous two approaches which require n * (n-1) /2 bidirectional systems by reducing the number of systems by a factor of (n-1)/2 for n language pairs intertranslation.

Though it seems promising at the first glance, extracting the Interlingua (manipulating the meaning) and the choice of the representation of the meaning are among the difficulties in this approach.

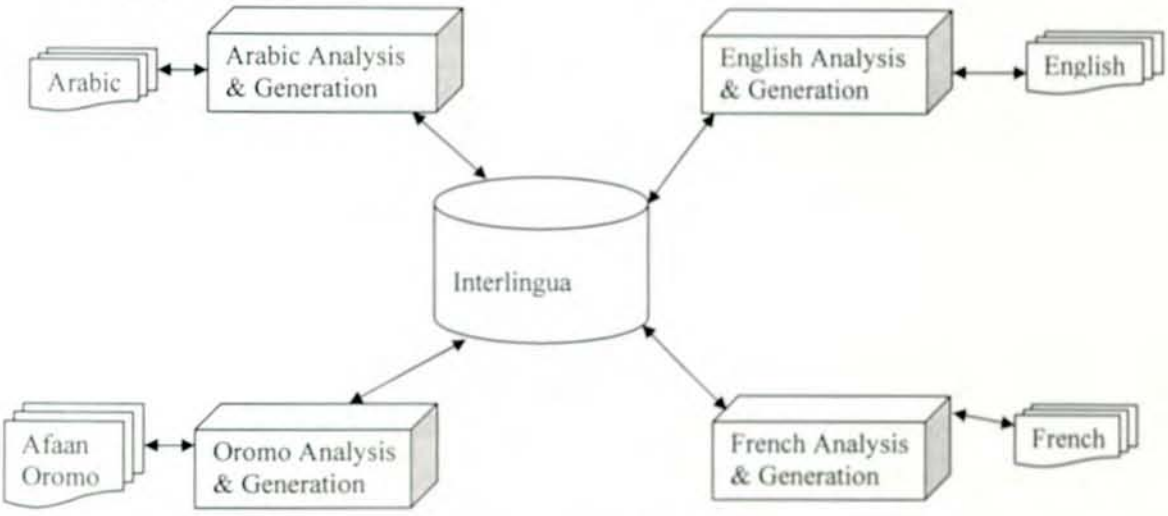


Figure 3.5. Translation using Interlingua approach

Vauquois triangle shown in Figure 3.6 below is a common way to visually present the above three approaches (Jurafsky and Martin, 2008). 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 interlingual approaches. 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).

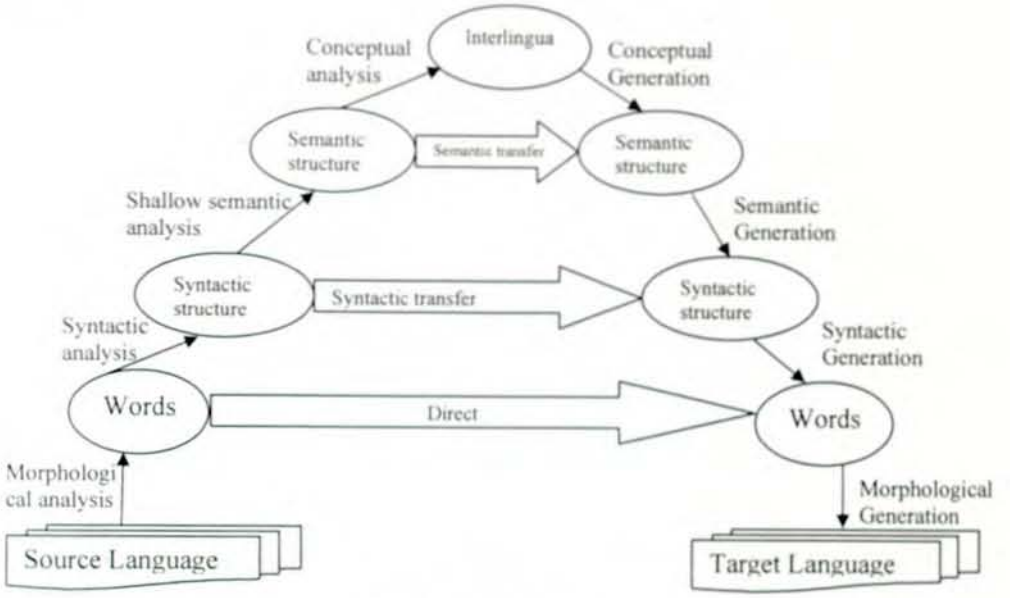


Figure 3.6 The Vauquois Triangle

3.4.5. Corpus-based Approaches

As of 1990's, MT research moved from classical rule-based approach to empirical or corpus-based systems. Empirical systems are data-driven as opposed to rule-driven. Traditional (rule-based) translation methods are somewhat difficult to construct, as they basically involve hardcoding the idiosyncrasies of both the source and target languages. But through the work of human translators, large parallel corpora have become available. Therefore, it makes sense to use these large bodies of already

existing translations to capture the knowledge contained in them. Hence, corpus-based MT is based fundamentally on the principle of using existing translations as a prime source of information for the production of new ones (i.e., it believes in the fact that large amounts of data contain essential knowledge for making a functional system) (Russell and Norvig, 2003; Jurafsky and Martin, 2008; Arnold et. al., 1994). In spite of the rule-based models that require explicit linguistic knowledge, data-driven ones rectify the lack of such knowledge in such a way that the knowledge can be retrieved and used automatically. Corpus-based approach has two varieties, namely, Example-Based Machine Translation (EBMT), and Statistical Machine Translation (SMT). The following subsections briefly explain these approaches.

3.4.5.1. Example-based MT (EBMT)

Example-based machine translation (EBMT) (also known as Translation by Analogy) approach (Arnold et. al., 1994) is often characterized by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by matching against stored example translations and can be viewed as an implementation of case-based reasoning approach of machine learning. For example, linguistic knowledge about word order, agreement, etc. is captured automatically from examples. It relies on large corpora and tries somewhat to reject traditional linguistic notions like part of speech and morphology (although this does not restrict them entirely from using the said notions to improve their output). EBMT systems are attractive in that they require a minimum of prior knowledge and are therefore quickly adaptable to many language pairs.

According to Arnold et. al. (1994), 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 good data and the quality increases as the examples become more and more complete (Arnold et.

al., 1994). 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). To see how EBMT works, consider the example of translating the sentence,

"The blue cat chased the brown mouse"

Using the corpus comprising the following two sentences pairs:

English	Afaan Oromoo
E1: The blue cat ate meat.	O1: Adurren cuquliisni foon nyaatte.
E2: The woman chased the brown mouse.	O2: Nadhittiin hantuuta magaala reebde.

During translation, parts of the sentence to be translated are matched with the sentences in the corpus. Here, The blue cat matches exactly with the underlined part in O1. Likewise, chased the brown mouse is matched exactly with the underlined part in O2. So, the two sentence fragments are taken and combined appropriately to get the meaning:

Adurren cuquliisni hantuuta magaala reebde.

Thus, EBMT entails the following three steps (Ramanathan, 2005):

1. Matching fragments against the parallel corpus,
2. Adapting the matched fragments to the target language, and,
3. Recombining these translated fragments appropriately.

However, problem arises when one has a number of different examples each of which matches part of the string, but where the parts they match overlap, and/or do not cover the whole string (Arnold et. al., 1994).

3.4.5.2. Statistical Machine Translation

The approach can be thought of as trying to apply the techniques which have been highly successful in Speech Recognition (Brown et. al., 1990, 1993) to MT. Though the details require a reasonable amount of statistical sophistication, the basic idea of the statistical models can be grasped quite simply. SMT is like EBMT except that it has probabilities (Koehn, 2003). The next chapter describes SMT in more detail since it is used in this research.

3.5. Challenges in Machine Translation

Different factors contribute to the difficulty of machine translation to achieve a high quality output. Language similarity and difference has the highest contribution. Language similarity and difference can be categorized into two: systematic differences which can be modeled in a general way, and idiosyncratic & lexical differences that must be dealt with one by one (Jurafsky and Martin, 2008). Syntactically, unlike English, Afaan Oromoo is morphologically rich. Afaan Oromoo has agglutinative morphology while English has fussional morphology, in general. Unlike English, Afaan Oromoo is subject prodrop langauge. For example, "deeme" (meaning "went") can be a sentence by dropping either "inni" (meaning "he") or "ani" (meaning "I"). Therefore, "I went" and "He went" can be translated to "Deeme", "Inni deeme" or "Ani deeme". (Detail of English Versus Afaan Oromoo is given in Chapter two.)

CHAPTER FOUR

STATISTICAL MACHINE TRANSLATION

4.1. Introduction

In this chapter, statistical machine translation is reviewed with the objective of getting insight into the different components, algorithms, and tools that are used in statistical machine translation. It starts with the general introduction of SMT and keeps on explaining the components and the algorithms.

4.2. Noisy – channel Model

Since the research of the last decade of the 1980's at IBM (Brown et. al. 1990), SMT attracted many research endeavors in the area of machine translation. This research is also the result of that attraction.

Statistical machine translation is based on the notion of noisy-channel model (Brown et.al., 1990) to combine language model (discussed in section 4.3) and translation model (discussed in section 4.4) as shown in Figure 4.1.

Given a source string S to be translated into a target string T (i.e., during the translation task), the model considers S as the target of a communication channel, and its translation T as the source of the channel. Thus, viewing every Target string as a possible source for each Source string, the machine translation task is to recover the source from the target by assigning a probability Pr(T|S) to each pair of sentences(T,S), and seeking the particular T which maximizes Pr(T|S).

Beware! Do not get confused at this point! It is not a trivial concept, there are two tasks namely Modeling Task which is the mathematical model and Translation Task which is the application of the model. The two tasks consider the source language and target language differently. For example, in this system (English – Afaan

Oromoo), Afaan Oromoo is considered the source language for modeling task whereas the source language for translation task is English as the following quote from the SMT textbook by Koehn (2006) states, .

> *"This may create a lot of confusion, since the concept of what constitutes the source language differs between the mathematics of the model and the actual application."* (Koehn, 2006)

This concept is not as clear as it sounds at first glance. After all, it is what kept SMT dormant for about a decade until Knight (1999a) shade light on it. If you don't understand this, do not proceed to other sections in this chapter as it will get worse and worse for you. If you are brave enough and want to know its detail, refer to its mathematical foundation in Brown et. al. (1993)

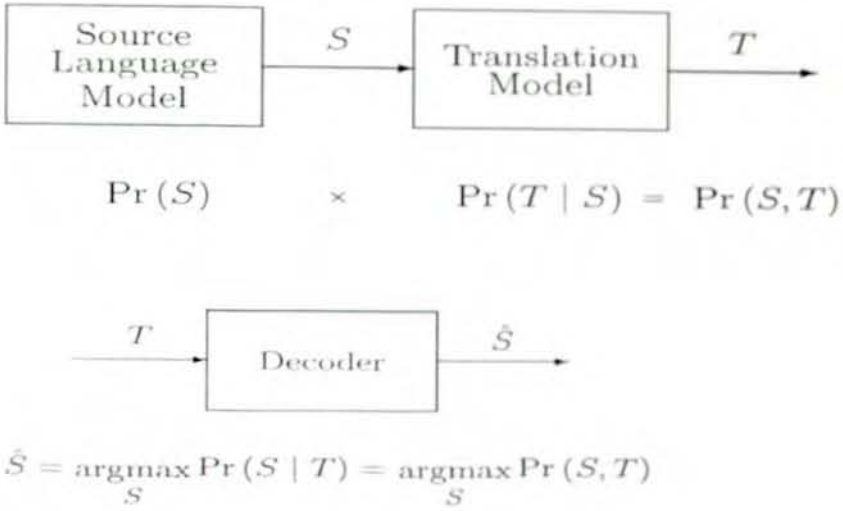


Figure 4.1 Source-channel modeling (Brown et.al, 1990)

Figure 4.1 depicts two key notions in SMT. These two key notions involved in SMT are the language model that models the fluency of the translated string and the translation model that models the adequacy of the translated string (Russell and Norvig, 2003). After modeling the translation fluency and adequacy, the best target

language document that matches in meaning to the source document is obtained by decoding. Decoding is a search problem that makes use of the models built and the source document to come up with the best target document. Once we have the target document, we can evaluate how good our system performed as compared to the human translation of the given document. This is the evaluation of the system. The following subsections deal with language modeling, translation modeling, decoding and evaluation.

4.3. Language Modeling

The language model provides the probabilities for strings of words (in fact sentences), which can be denoted by P(S) (for a source sentence S) and P(T) (for any given target sentence T) (Brown et. al., 1990,1993). Intuitively, P(S) is the probability of a string of source words occurring in S, likewise, P(T) is the probability of a string of target words occurring in T. The goal of language modeling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible. It is committed to handling the word reordering in the language in question. It is used to get how fluent the translated text is in the target language, i.e., the sequence of string of words that has the highest probability is assumed to be the most fluent sentence in the target language.

There are different techniques for statistical language modeling including n-gram language modeling (Jurafsky and Martin, 2008), maximum entropy language modeling (Berger et.al. 1996), and whole sentence exponential modeling (Rosenfeld et. al, 2001).

From these techniques, N-gram model is the most widely used statistical language model. It has been used successfully in speech recognition, spell checking, par-of-speech tagging, machine translation and other tasks where language modeling is

required (Ramanathan, 2005; Lopez, 2008). Since N-gram technique is used in this research, the following paragraphs explain it in detail

It is useful to use n-grams in language modeling (rather than using the whole sentence) for two reasons: one is that it keeps things to manageable limits and the other is that the probability converges to zero if we take all the words in a long sentence. In order to calculate these source language probabilities (producing the source language model by estimating the parameters), a large amount of monolingual data of the target language is required, since the validity, usefulness or accuracy of the model will depend mainly on the size of the corpus.

Using n-gram technique, the probability of a sentence S containing the word sequence $w_1w_2...w_n$ can be expressed, without loss of generality, by using the chain rule as

$$p(S) = p(w1) \times p(w2|w1) \times p(w3|w1, w2) \times ... \times p(wn|w1, w2, ...wn-1)$$

$$= \prod_{i=1}^{n} (\mathrm{p}(\mathrm{wi}|\mathrm{w1}..\mathrm{wi}-1)$$

From this equation, it is clear that the probability of the sentence S is simply the product of many small probabilities, each of which corresponds to the probability of a single word. Intuitively, the conditional probability $P(w_2|w_1)$ is the probability that w_2 will occur, given that w_1 has occurred. For example, the probability that *am* and *are* occur in a text might be approximately the same, but the probability of *am* occurring after *I* is quite high, while that of *are* occurring after *I* is much lower.

Since the distribution $p(w_n|w_1...w_{n-1})$ that is assigned to the last word of the sentence contains nearly as many terms as p(s) itself, finding $p(w_n|w_1,w_2,...w_{n-1})$ will be as difficult as finding p(s). To simplify this, the idea of conditional independence (Lopez, 2008) is introduced. The conditional independence that will be assumed here is that

the probability of word w_i is conditionally independent of all but the preceding n-1 words, whatever value of n is taken. This conditional independence assumption is the n-gram approximation. For example, if the value of n is 2 (i.e. bigram model), the probability of a word w_i is conditionally dependent on the preceding word (i.e., 2-1) only. Likewise, if the value of n is 3 (i.e. trigram model), the probability of a word w_i is conditionally dependent on the preceding 2 words (i.e., 3-1).

From the given corpus, using maximum likelihood estimate, a bigram model is given by:

$$p(w2|w1) = \frac{count(w1w2)}{count(w1)}$$

Similarly, in a trigram model, $p(w3|w2,w1) = \frac{count(w1w2w3)}{count(w1w2)}$

And so on, where w_1w_2 *and* $w_1w_2w_3$ are word sequences in a string.

For example, in a bigram language model (n=2), the probability of the sentence "Mucaan guddaan deeme." is approximated as

P("Mucaan") x P("guddaan"|"Mucaan") x P("deeme"|"guddaan") x P(".|"deeme")

whereas in a trigram language model (n=3), the approximation is

P(Mucaan)P(guddaan|Mucaan)P(deeme|Mucaan guddaan)P(.|guddaan deeme)

The most commonly used statistical language modeling software tools available to statistical language modeling community include: CMU-Cambridge Statistical Language Modeling toolkit (Clarkson and Rosenfeld,1997), SRI Language Modeling Toolkit (Stolcke, 2002), N-gram Stat (Banerjee and Pedersen, 2003), and Trigger Toolkit (Berger, 1997).

The choice of a tool from the available ones requires one to deal with some criteria like compatibility with other subcomponents to be used, and update and maintenance or its currency. From these tools, the researcher believes that SRI Language Modeling Toolkit is more appropriate for the reason that it is compatible with moses (Koehn et. al, 2007) decoder that will be used in this research. Furthermore, it is regularly updated.

4.4. Translation Modeling

The translation model, on the other hand, provides us with probability P(T|S) (which is read as "probability of T given S") is the conditional probability that a target sentence T will occur in a target text which translates a text containing the source sentence S. As shown in the noise-channel modeling diagram above (Figure 4.1), the product of this and the probability of S itself, that is P(S) * P(T|S) gives the probability of source-target pairs of sentences co-occurring, written P(S,T) which is given by,

$$P(S,T) = P(S) \times P(T|S)$$

In order to come up with the translation model, constructing word alignment, which will be discussed below, is a crucial task.

Word Alignment

The fundamental input to the translation modeling is the alignment. Word alignment is a mapping between the words in the source sentence and the words in the target sentence. The alignment A for source sentence S = $s_1 \ldots s_l$ *and* target sentence T = $t_1 \ldots t_m$ is given by,

$A = \{a_1, a_2, \ldots a_m\}$ where $a_j \in \{0,1,\ldots l\}$

Given *l* and *m* are lengths of source and target sentences respectively, alignment *A* indicates which source word generates each target word in the given sentence.

For example, given, S = "The cat ate meat." and T = "Adurreen foon nyaatte.", one best alignment from 6^4 or 1296 possible alignments is A = {2,4,3,5}, i.e, the first word in the target sentence is generated by the second word in the source sentence; the second word in the target sentence is generated by the fourth word in the source sentence and so on as pictorially visualized below.

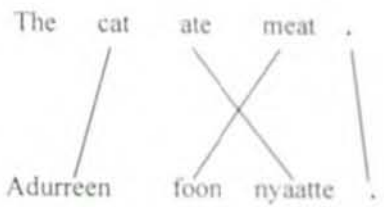


Figure 4.2 The best out of 1296 possible alignments

The IBM team proposed a series of five models (IBM Model 1, IBM Model 2, IBM Model 3, IBM Model 4, and IBM Model 5) (Brown et al, 1993) for alignment in translation modeling. Below, reviewed are the models with more detail of IBM Model 1 as the major task of estimating lexical translation is done in this model.

IBM Model 1

IBM Model 1 (Brown et. al., 1993) is the simplest model among the models that the IBM team proposed to estimate lexical translation parameter T such that $T(t_j \mid s_{a_j})$ is translation probability of target word t_j given the source word s_j that generated it in the given alignment where a_j is $\in$ A. The general generative story of IBM Model 1 is about how we generate target sentence from a source sentence S = $s_1, s_2, \ldots s_l$. The IBM Model 1 generative story is:

- Given S – the source sentence:
- Let T be the target sentence
- Pick $m = |T|$, where all lengths m are equally probable

- Pick alignment A with probability $P(A|S) = 1/(l+1)^m$, since all alignments are equally likely given l and m, where $l = |S|$
- Pick $t_1 \ldots t_m$ in T with probability $P(T \mid A,S) = \prod_{j=1}^{m} T(t_j \mid s_{a_j})$ where $T(t_j \mid s_{aj})$ is the translation probability of target word t_j given the source word it is aligned to.

Example,

S= "tall boy"

Pick m = |T|=2

T = "t_1t_2"

From the corpus,

(S,T) = <"tall boy", "mucaa dheeraa">

Pick A = {2,1} with probability $1/(l+1)^m$

Pick t1 = "mucaa" with probability P("mucaa"|"boy)

Pick f2 = "dheeraa" with probability P("dheeraa"|"tall")

Since the model for $P(T|S)$ contains the parameter $T(T_j \mid S_{aj})$, we first need to estimate $T(t_j \mid s_{aj})$ from the data. If we have the data and the alignments A, along with $P(A|T,S)$, then we could estimate $T(t_j \mid s_{aj})$ using expected counts as follows:

$$T(t_j \mid s_{a_j}) = \frac{Count(t_j, s_{a_j})}{\sum_{t_{j'}} Count(t_{j'}, s_{a_j})}$$

But we don't have P(A|T,S). To estimate P(A|T,S), we use P(A|T,S) = P(A,T|S) / P(T|S)

But $P(T \mid S) = \sum_{A \in A} P(A,T \mid S)$

So we need to compute *P(A, T|S)* which is given by the Model 1 generative story:

$$P(A,T \mid S) = \frac{C}{(l+1)^m} * \prod_{j=1}^{m} T(t_j \mid s_{a_j})$$

Therefore, for the above example,

$$P(A|T,S) = P(T,A|S)/P(T|S)$$

$$= \frac{\frac{C}{3^2} * T(mucaa \mid boy) * T(dheeraa \mid tall)}{\sum_{A \in A} \frac{C}{3^2} * \prod_{j} T(t_j \mid s_{a_j})}$$

, where C is a normalization constant.

So, in order to estimate P(T|S), we first need to estimate the model parameter

$T(t_j \mid s_{aj})$; in order to compute $T(t_j \mid s_{aj})$, we need to estimate P(A|t,s); and in order to compute P(A|t,s), we need to estimate $T(tj \mid s_{aj})$. Here, we got into a chicken-egg problem where EM comes to play.

Expectation Maximization (EM)

Expectation maximization (also known as estimation maximization (Knight, 1999a) solves the chicken-egg problem above using the following steps (Jurafsky and Martins, 2008).

Given the training data set of pairs $< s_i, t_i >$, log likelihood of training data given model parameters is:

$$\sum_i \log P(T \mid S) = \sum_i \log \sum_{A \in A} P(A \mid s_i) * P(t_i \mid A, s_i)$$

To maximize log likelihood of training data, given model parameters, use EM where the hidden variable is the alignments A and the model parameters is the translation probabilities T. The following are the steps in EM.

- Initialize model parameters *T(T|S)*

- Calculate alignment probabilities $P(A|T,S)$ under current values of $T(T|S)$
- Calculate expected counts from alignment probabilities
- Re-estimate $T(T|S)$ from these expected counts
- Repeat until log likelihood of training data converges to a maximum

IBM Model 2

Model 1 does not worry about where the words appear in either of the strings. So, Model 2 builds on top of Model 1 to reorder the words in the target sentence. The Model 2 parameters are the lexical translation probability ($T(t_j \mid s_{a_j})$ which is equal to the translation probability of target word t_j given source word s_{a_j} that generated it) which is estimated in Model 1 and the distortion probability (d(j|i,l,m) which is equal to the distortion probability, or probability that t_i is aligned to s_i, given l and m) which will be estimated here. For example, d(2|3,5,6) means, the probability that t_2 is aligned to s_3 where |s| = 5 and |t| = 6.

IBM Model 3

Not all words in the source language map to exactly one word in the target language. Model 3 adds the fertility probability ($n(s_j)$ which is equal to the likelihood of each source word translated to one word, two words, three words, and so on) on top of Model 2 parameters. Model 3 has the following parameters:

- $T(t_j \mid s_{aj})$ = translation probability of target word t_j given source word s_{aj} that generated it
- $d(j|i,l,m)$ = distortion probability, i.e., probability of position t_j, given its alignment to s_i, l, and m
- $n(s_i)$ = fertility of word s_i, i.e., number of target words aligned to s_i
- p_1 = probability of generating a target word by alignment with the NULL source word. This is also known as spurious word probability (Knight, 1999a)

IBM Model 4

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

IBM Model 5

According to Brown et. el (1993), Model 5 removes the deficiencies of the previous models. For example, Model 4 can stack several words on top of one another. It can also place words before the first position or beyond the last position in the target string. Therefore, Model 5 fixes deficiencies like this one that the previous models have not handled.

Software tools available for alignment and translation modeling include: GIZA++ (Och and Ney, 2003), Twente (Hiemstra, 1998), and K-vec++ (Pedersen and Varma, 2001), and moses (Koehn et. al, 2007).

4.5. Decoding

Once the translation model and language model is built, decoding tries to find the highest probability translation of a sentence or phrase among the exponential number of choices for a new input source sentence. This is the essence of the approach to SMT, although the procedure is itself slightly more complicated in involving search through possible source language sentences for the one which maximizes Pr(S) * Pr(T|S), translation being essentially viewed as the problem of finding target sentence T that is most probable, given the source sentence S. Put mathematically,

$$P(S,T) = argmax(P(S|T))$$

$$= argmax_s\left(\frac{P(S)P(T|S)}{P(T)}\right)$$

$$= argmax_s(P(S)(P(T|S))$$

Then, one just needs to choose S that maximizes the product of Pr(S) and Pr(T|S) as Pr(T) is not dependent on S.

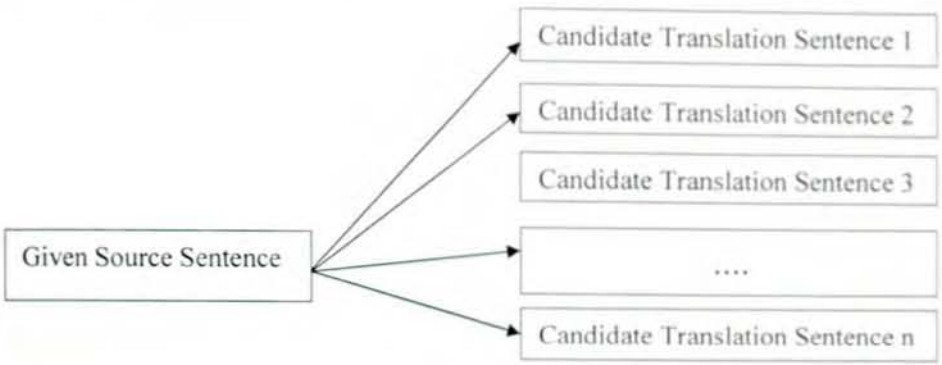


Figure 4.3 Decoding

Decoding or finding the sentence that maximizes the translation and language model probabilities is a search problem. Decoders in MT are based on best-first search (Jurafsky and Martins, 2008). A* search (a variant of best-first search) was first implemented for word-based SMT by IBM (Brown et. al.,1995). However, the currently publicly available moses (Koehn et. al., 2007) is implemented using beam search for phrase-based decoding.

Searching in SMT involves starting with a state with empty hypothesis. Each state is associated with the sum of current cost (which is the total probability of the phrases that have been translated so far in the hypothesis) and the future cost (the estimate of the cost of translating the remaining words in the target sentence). The cost is the product of the translation, distortion, and language model probabilities.

Decoding process suggests that the state space of possible translations be searched which leads to expanding the entire search space in which case the problem will be NP-complete (Knight, 1999b). Therefore, in order not to come up with a graph that is too large to fit into memory, pruning of high-cost states and keeping only the most promising states is very important.

Software tools available for decoding that are widely used include Pharao (Koehn, 2004), and moses (Koehn et. al., 2007).

4.6. Evaluation

Evaluating the quality of a translation is essential, although it is extremely subjective. In MT, evaluation is done along two dimensions: fidelity and fluency. It can be done manually by human raters or automatically.

Evaluation Using Human Raters

The most accurate evaluations use human raters to evaluate each translation along each dimension. For example, clarity, readability, and the degree of naturalness of MT output is measured along the dimension of fluency.

Fidelity, on the other hand, is measured by adequacy and informativeness. The adequacy of a translation is whether it contains the information that existed in the source. Informativeness of a translation is a task-based evaluation of whether there is sufficient information in the MT output to perform some task. Both fluency and fidelity can be measured by giving raters a scale that show the degree of fluency or fidelity so that they can judge the MT output and rank it in the given scale.

The combined evaluation of fidelity and fluency to measure the overall quality of translation can be done using edit-distance (edit cost of post-editing the MT output).

It can take different form like the number of key strokes, time taken to post-edit, or number of words to correct.

Automatic Evaluation

Despite the fact that human evaluation is the most reliable, it is very time consuming and boring if done repeatedly. Therefore, with some compromise of quality, automatic evaluation that has high correlation with human evaluation is better to use as it will run frequently to go for system improvements quickly.

Automatic evaluation methods include BLEU (Papineni et. al, 2002), NIST (Doddington, 2002), Precision and Recall (Turian et. al., 2003), and METEOR (Banerjee and Lavie, 2005).

As BLEU is used in this research, it is important to have a closer look at it. It works by ranking each MT output by a weighted average of n-gram precision against a set of human translations as references. It measures what percentage of n-grams in the MT output is also found in a human translation. For example, if an MT output sentence has 10 words and 4 of which occurred in the reference translation, using unigram matching, this system has 4/10 precision. This is not good as it favors output sentences with repeated words. If, in the above example, the output sentence has 3 of them the same, it is not fair to give it a precision score of 4/10. To overcome this, BLEU uses modified n-gram precision (p_n) metric such that the count of each candidate word is divided by the maximum number of times a word is used in any single reference translation.

When computing the score of the whole testset in BLEU, first the n-gram matches for each sentence is computed, then the count is clipped (penalized by dividing it with the maximum reference count) and added together for all candidates and divided by

the total number of candidate n-grams in the testset as shown in the following formula.

$$p_n = \frac{\sum_{c \in \{candidates\}} \sum_{n-gram \in c} count_{clip}(n-gram)}{\sum_{c' \in \{candidates\}} \sum_{n-gram \in c'} count_{clip}(n-gram')}$$

The other problem is evaluating output sentences that have fewer words than the reference translation that will abnormally occur in many reference translations. BLEU includes a brevity penalty or length penalty over the whole corpus to overcome this problem. Let c be the total length of the candidate translation for the corpus. The effective reference length r for that corpus is computed by summing, for each candidate sentence, the length of the best matches. The brevity penalty (BP) is then an exponential in r/c. Put mathematically,

$$BP = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$

Therefore, the computation of BLEU is given by,

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} \log p_n\right)$$

4.7. Related Works

There are many researches on machine translation since its inception. Prior to 1990, researches on rule-based approach dominate the field of MT. In the case of Ethiopian languages, an MT research for English - Amharic was conducted in 2004 at AAU using rule-based approach (Yihenew, 2004) and number of researches that can be input to English – Amharic MT have been conducted by Carl and Sisay (2003), Sisay and Haller (2003a) and Sisay and Haller (2003b).

When it comes to the statistical approach, Candid (Brown et. al., 1990) was the first major prototype that was developed as a result of the groundbreaking research on

MT by a team of researchers in IBM. This team used the Canadian Hansard corpus and the IBM statistical models (Model 1, Model 2, Model 3, Model 4, and Model 5) drawn from the algorithms of speech recognition (Brown et. al., 1990, 1993) to build Candid. These IBM models are considered the first generation of SMT as they are primarily word-based models (Brown et. al, 1990) as opposed to phrase-based models (Koehn et. al, 2003).

Since then, progresses have been achieved in alignment (from word-based (Brown et. al.,1990;1993) to phrase-based (Koehn et. al., 2003; Och and Ney, 2004) to syntax-based (Zhang et. al., 2007; Charniak et. al., 2003)); in searching or decoding (from best first or pure A* (Brown et. al., 1995) to beam search (Koehn et. al., 2003)); and in corpus development (the French-English Canadian Hansard, the German-English VerbiMobil, and the multilingual Europarl).

SMT researches have been conducted for many language pairs, out of which Euromatrix (Koehn, 2005) project (Which is conducted for the intertranslation of 11 languages yielding 110 systems) is the largest. Moreover, SMT researches have been conducted for a number of language pairs, in addition to the ones included in the Euromatrix project. Chinese-English SMT (Su, 2004), Estonian-English SMT (Fishel et. al, 2007), and Tamil-English (Germann, 2001) are among the language pairs for which SMT approach has been researched on.

CHAPTER FIVE

ENGLISH – AFAAN OROMOO SMT SYSTEM

5.1. Introduction

In this chapter, the proposed English – Afaan Oromoo SMT system will thoroughly be discussed. In the next sections, the discussion on the architecture of the system that includes the schematic representation of the system with its software components, inputs and outputs will be made.

In addition, the experimental setup that includes the hardware environment on which the experimentation is conducted, the software tools used for each component of the system, and the data used for the experimentation of the research will be discussed. Finally, the process of the experimentation and the results as well as the analysis of the results will be explained.

5.2. Architecture of the System

The architecture of the English – Afaan Oromoo SMT system that is shown diagrammatically in Figure 5.1 includes the four basic components of Statistical Machine Translation discussed in CHAPTER FOUR: Language Modeling, Translation Modeling, Decoding and Evaluation.

The Language Modeling component takes the monolingual corpus and produces the language model for the target language. The Translation Modeling component takes the part of the bilingual corpus 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 the system output and the reference translation of the input to produce the metric that compares the system output and the reference translation.

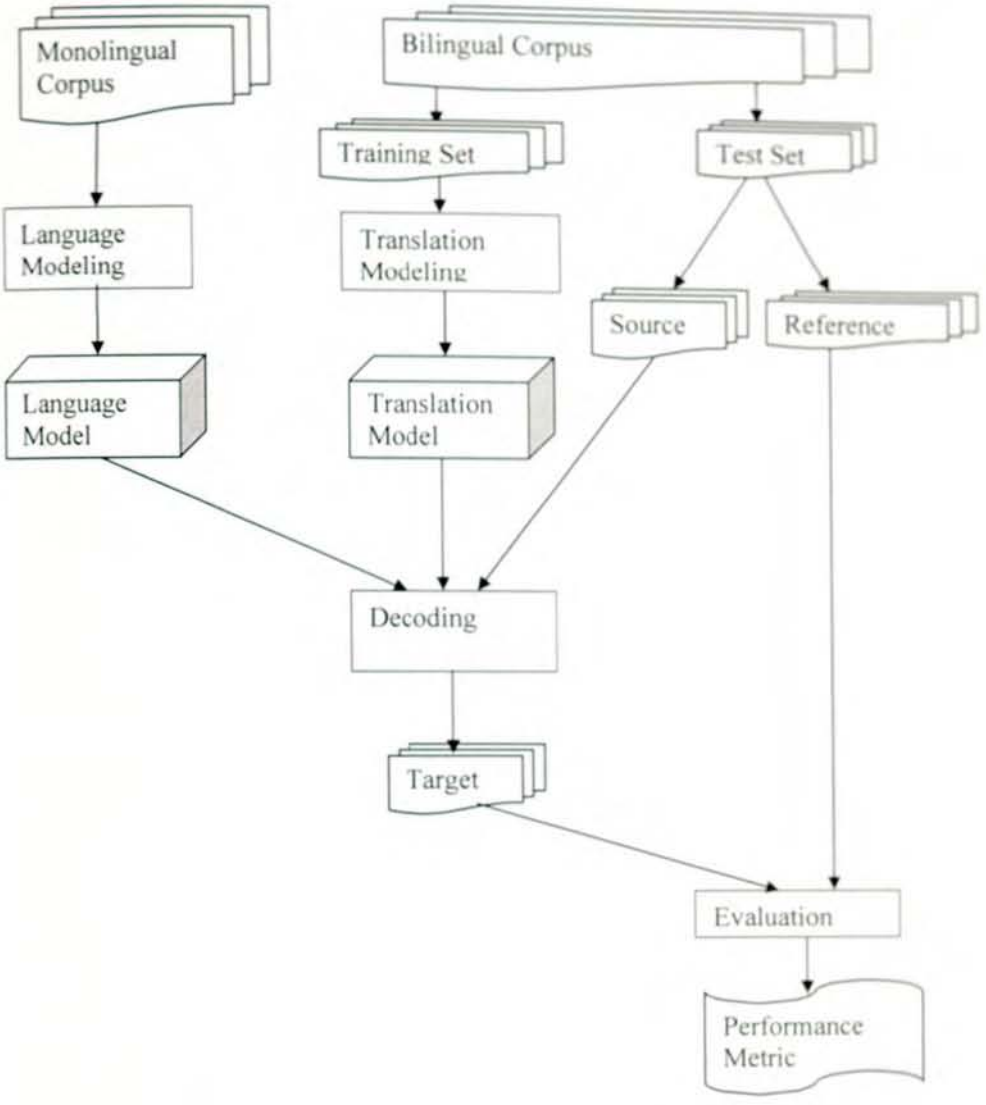


Figure 5.1 Architecture of the English – Afaan Oromoo SMT System

In Figure 5.1 above, the SMT components are represented by rectangle, the models by cube, the data by pile of sheet, the data flow by arrows, and performance report by a sheet.

5.3. Experimental Setup

5.3.1. Data

To find P(S,T) discussed in chapter four, there are two tasks to be accomplished both of which require large amount of data. One task that requires large amounts of monolingual data is, then, to find out the probability of a source string (or sentence) occurring (i.e., P(S)) which has been taken care of by the language modeling discussed in section 4.3.

The second task requiring large amounts of data is specifying the parameters of the translation model, which requires a large bilingual aligned corpus. There are rather few such resources, however. The research group at IBM which has been mainly responsible for developing this approach had access to about three million sentence pairs from the Canadian Hansard (French-English) — the official record of proceedings in the Canadian Parliament, from which they have developed a sentence-aligned corpus, where each source sentence is paired with its translation in the target language (Brown et. al. 1990). Not only is the corpus resources availability but also its usefulness depends very much on the state in which it is available to the researcher. Corpus clean-up and especially the correction of errors such as those arising from transencoding is a time-consuming and expensive task, and one can argue that it detracts from the 'purity' of the data as cleaning by modification may result in 'artificial corpus'. The collection, processing and organization of the data used in this research are discussed as follows.

5.3.1.1. Collection of the data

The researcher found digitally available Afaan Oromoo versions of some chapters of the Bible and some spiritual manuscripts for which the English counterparts were explored on the web. While religious material may not contain the day-to-day language words, it has the advantage of being inherently aligned on the verse level,

facilitating further sentence alignment. Another typical bilingual text is the United Nation's Declaration of Human Rights, which is available in many of the world's languages, including Afaan Oromoo. Fortunately enough, this text was manually sentence-aligned. The Kenyan Refugee Act and the Ethiopian Constitution, which are manually sentence aligned, are also among the most important documents the researcher found on the web. Moreover, the researcher was able to locate other documents on the web that have medical content.

In addition to the ones obtained from the web, the Council of the Oromia Regional State (Caffee Oromiyaa) provided the researcher with the legislations and proclamations that have been promulgated in the regional state in the past years (i.e., 1991 – 2007). The council provided the researcher not only the proclamations but also the regional constitution of Oromia Regional State.

5.3.1.2. Preliminary Preparation

The collected raw files were in different formats and encodings. Some of them were even in hardcopies. After having typed the hard copy corpus and merging it with what has already been collected, manipulation of the data to put it into uniform format and encoding was necessary. Some of the documents were doc files and html files which were pretty easy for format trasconversion and trasencoding. Some of the documents were presented in colorful brochures in PDF format which made the result of the automatic text extraction tool less useful without manual postprocessing. The other documents were in Adobe PageMaker format which demanded a lot of time getting the text content looselessly. Therefore, a major difficulty in the preparation task was extracting the data from these documents. All of the data in the 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.

5.3.1.3. Size of the Data

The data collected for language modeling and translation modeling – the monolingual and bilingual corpus respectively, is described as follows.

- **Monolingual Corpus**

The monolingual corpus that is necessary for training the fluency of the target language is collected from different sources. Professor Kevin Scannel (http://borel.slu.edu/crubadan/) provided the researcher some Afaan Oromoo corpus which his crawler has collected from the web. With the objective of increasing the size of the monolingual corpus that is used to estimate the fluency of a sentence in the language, half (i.e., Afaan Oromoo part) of the bilingual corpus was also included in this set. The monolingual corpus contains 62,300 sentences (1,024,156 words).

- **Bilingual Corpus**

The corpus currently consists of 20,000 sentences (300,000 words in each language), although this is not comparable to the resources available to Indo-European language pairs, such as the Hansard corpus of size 2.87 million sentence pairs (Roukos et al., 1997) and EUROPARL corpus of size 1.3 Million sentence pairs on average (Koehn, 2005). The following table (Table 5.1) summarizes the amount of data used in this research.

Units	Bilingual		Monolingual
	English	Afaan Oromoo	Afaan Oromoo
Sentences	21, 085	20848	62,300
Words	384,881	308,051	1,024,156

Table 5. 1 Summary of the size of the total data used in the research

5.3.1.4. Organization of the Data

The data is organized into training and testing data in the proportion of 9:1 because of the size of the data. The researcher believes in the fact that the better the system

learns the better it answers. That is why it is planned to train the system with 90% of the data and test with the remaining 10%. The data is split randomly into two to make training, and test set.

- **Training set**

The training data consists of 19,058 sentences (354,978 words) of English and 18,699 sentences (280,880 words) of Afaan Oromoo which were used for training the translation model.

- **Test Set**

The test set is the part of the bilingual corpus that whose source text is given to the system and whose target text (also known as reference translation) is compared against the output of the system. The test set contains 2,027 sentences (29,903 words) of English and 2,149 sentences (27,171 words) of Afaan Oromo.

In addition to the preparation of the data discussed section 5.3.1.2 above, preparing the test set requires wrapping up the source and the reference text with SGML markup. This is done automatically using a script written using python for this purpose. As the automatic script for SGML marking up is not accurate enough due to the inaccuracy of the sentence aligner, the source and the reference translation are edited manually for the proper markup.

In order to be in the right format expected by the evaluation script, the reference translation file is wrapped in the SGML markup as follows:

```
<refset setid="enom" srclang="en" trglang="om">
      <doc docid="enom" sysid="ref">
            <seg id=#senteceID> line of sentence here </seg>
      </doc>
</refset>
```

Likewise, the source file should also be wrapped up with SGML markup showing the description of the system, the document and the segments as follows:

```
<srcset setid = "enom" srclang="en">
      <DOC docid="enom" sysid="input">
            <seg id=#> #line of sentence here </seg>
      </DOC>
</srcset>
```

5.3.2. Software Tools Used

The training of the system was done is 32 bit linux machine as an operating system platform. Moreover, there are different scripts and software tools for each component of the SMT system whose blueprint is shown in section 5.2 above.

- **Preprocessing**

Preparation of the monolingual and the parallel corpus data was done using different scripts written for this purpose (available at http://www.statmt.org). Those scripts which have been written for preprocessing of corpus before it is sent to the actual training have been customized to handle peculiar behaviors of Afaan Oromoo like the apostrophe. Sentence aligning, tokenization, lowercasing and truncating long sentences that take the alignment to be out of optimality were done by those scripts.

- **Language modeling**

The widely used language modeling tool SRILM toolkit (Stolcke, 2002) was used for language modeling because the moses MT system has a support for SRILM as a language modeling tool.

- **Word Alignment**

For word-alignment, the state-of-the-art method is GIZA++ (Och and Ney, 2003), which implements the word alignment methods IBM1 to IBM5. While this method has a strong Indo-European bias, it is nevertheless interesting to see how far it can be used in Cushitic languages like Afaan Oromoo with the default approach used in statistical MT.

- **Decoding**

Decoding is done using moses (Koehn et. al, 2007) which is an SMT system that is used to train translation models to produce phrase tables. Moses decoder works using beam search algorithm (Koehn, 2004) for searching the best among the candidate translations.

- **Postprocessing**

Now that the data is prepared, the model is built and the system is given the test data and produced the output, the obtained output text must be postprocessed to be evaluated against the reference. The tasks in the postprocessing are: recasing (bringing the lowercased text to its proper cased equivalent), detokenizing (bringing the tokenized text to its human readable equivalent), and wrapping the output in SGML markup in order to compare the output with the marked up reference text segment for segment where a segment is a single sentence.

Once the recaser is trained on the already available cased corpus, it can guess what case the words should have in a sentence and creates the recased output file.

The next task is detokenizing the recased output. The detokenizer removes the extra spaces between strings and punctuation marks that were inserted to treat the punctuation marks and strings separately during training and testing.

Next is wrapping the output in SGML markup to make it ready for BLEU.

Different scripts are available at http://www.statmt.org that can do the recasing, detokenizing and wrapping with SGML and have been customized to support Afaan Oromoo.

- **Evaluation**

Evaluation is done using the BLEU (Papineni et al., 2002) scoring tool. Using a reference translation prepared manually from the parallel corpus, the translation quality of the system output which was translated can be evaluated. This is done by a script 'mteval-v11b.pl' which is available at http://www.statmt.org website.

5.3.3. Hardware Environment

The system was trained and tested on the powerful computational infrastructure which is dedicated for resource intensive researches such as machine translation in the Department of Computational Linguistics and Phonetics at Saarland University, Germany.

Access to sixteen clusters of computers suitable for CPU intensive tasks that require little memory and one computer which is suitable for memory intensive tasks was given to me. Since mine is of the second category, I used the system that is suitable for memory intensive task. Hence, the experiment of this research was conducted on a server that has 8 AMD Opteron 8220 dual core processors operating at 2.8GHz and 128GB RAM. However, since model building is the task that requires high computational resource, once the models are built, the system can be used on any computer without requiring as much memory as that in building the models.

5.4. Experiment and Analysis

5.4.1. Preprocessing

- **Tokenization**

When I try to inspect the output of the sentence-aligner for the Oromoo documents, I noticed that a word is split into three tokens (one to the left of the appostrophe, the appostrophe itself, and one to the right of the appostrophe). This is due to the fact that, the sentence-aligner does not consider a word having " ' " (apostrophe) as one word. So, I have to modify the preprocessor in such a way that it should not separate a word into three pieces, rather as a word having the appostrophe as a character.

In Afaan Oromoo, when this symbol comes at the end or beginning of a word, it is used as a single quote. Otherwise, it is used to represent a sound called 'Hudhaa' that should be dealt with at preprocessing. Here, if "'" (apostrophe) appears to be within a word, the tokenizer should not consider it to stand by itself rather it should keep the characters to left of it, itself and the character to the right of it as one token.

- **Preparing Abbreviations**

Though not exhaustive enough, list of the abbreviations for Afaan Oromoo that is used for tokenization and sentence alignment was prepared manually.

5.4.2. Building and Testing the System

In Figure 5.1, it is shown that the language modeling subsystem takes the monolingual corpus as input. For this research, 62,300 sentences (1,024,156 words) of monolingual corpus (including the half part of the bilingual corpus - the Afaan Oromoo part) was used to train the Language Modeling subsystem.

Bilingual corpus of 21,085 sentence (384,881 words) of English and 20,848 sentences (308,051 words) of Afaan Oromoo corpus was used to build the initial translation model of the system. From this, 90% or 19,058 sentences (354,978 words) of English corpus and 18,699 sentences (280,880 words) of Afaan Oromoo corpus has been used to train the system and the remaining 10% or 2,027 sentences

(29,903 words) of English and 2,149 sentences (27,171 words) of Afaan Oromo corpus has been used for testing the system.

5.4.3. Postprocessing

In postprocessing, the output of the system is recased (i.e., it is put in proper case of natural language), detokenized back to its original form, and wrapped into SGML markup to make it suit the evaluation script.

As it has been shown in section 5.3.1.4, the output file is also wrapped up in the SGML markup to fit the reference translation in structure for the evaluation script as follows.

```
<tstset setid="enom" srclang="en" trglang="om">
      <DOC docid= "enom" sysid="output">
            <seg id=#> #line of sentence here </seg>
      </DOC>
</tstset>
```

These tasks are among the most tiresome tasks in this research. Because, as data is piped from one process to another, the defect injected by one process will propagate to the next process making the final output difficult to match up with the initial one. For example, wrapping up the output document sentences into SGML requires the output to have equal number of sentences as in the reference translation. The system is not guaranteed to give us this if the source and the reference do not have equal number of sentences, even after automatic alignment of sentences. Therefore, some manual intervention of making the source sentences equal the reference translation to solve these discrepancies is made as it is inevitably necessary.

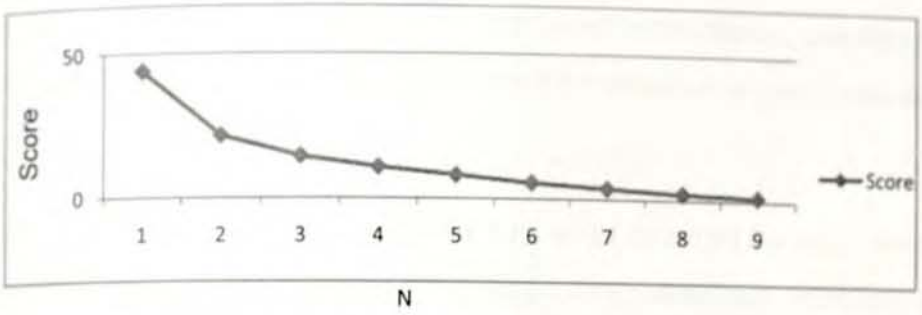


Figure 5.3 Individual N-gram Scoring

In addition to the limited size of the training corpus, the overall BLEU score of the system is attributed to the following major reasons:

- Availability of a single reference translation

A source sentence can have many possible translations. A very good MT output might look like one human translation, but very different from the other one. Thus, automatic evaluation metrics judge MT output by comparing with multiple human translations and taking the average. However, in this research, there is only one human translation for the given source sentences. As a result, a little deviation of the MT output from this single human translation will count against the score of this system which could not have been the case if there were many human translations. For example, I have seen an output sentence "Qooqa Oromoo" for the translation of source sentence "Oromo Language" which has a human translation of "Afaan Oromo" in the test data. From this example, one can easily note that even if the translation is correct, the automatic metrics will not consider it as good because the output translation is not found in this single reference translation.

- Domain of the test data

The test data is composed mainly of three major domains – legal, medical, and religious. The domain diversity of the test data has affected the BLEU score significantly. As the majority of my data is from the religious domain, upon evaluating

the output, I have investigated the bias introduced by the training data from religious domain. That is, the system performs better if it is tested on religious documents than documents from other domain.

When seen separately, the BLEU score for the test data from the legal, medical, and religious domains are 13.69%, 1.97%, and 21.72% respectively. From this, one can conclude that the BLEU score of the system is highly dependent on domain of the training and testing data.

5.4.5. An Attempt to Improve the Result

In an attempt to improve the result, the researcher tried to deal with the corpus size by bootstrapping as long as there is improvement in the BLUE score. Once the system was capable of translating English documents to Afaan Oromoo documents, iterative increase of the corpus size is possible by having translated English documents that have no Afaan Oromoo equivalent and then using them as parallel corpus. Two successive runs of the system (using 1000 English sentences in each run) and then retrain and retest showed deterioration of the BLEU score as shown in the following graph. Therefore, one can conclude that the initial corpus is not sufficient enough to train a system that produces good quality output that can be used in retraining the system.

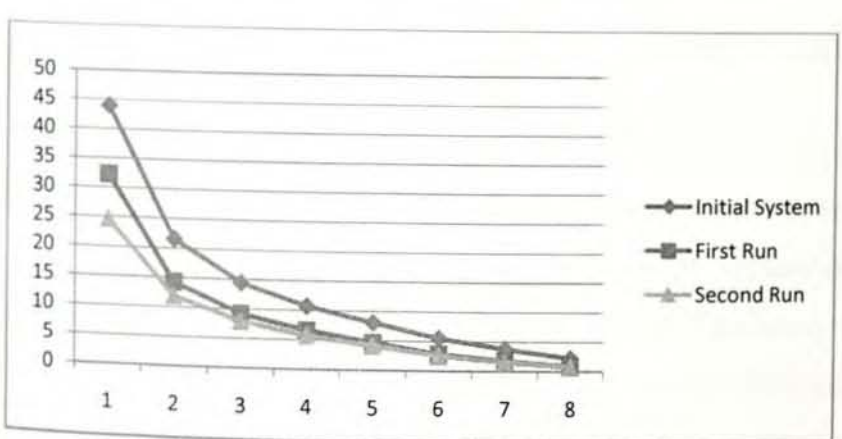


Figure 5.4 Result of training with unknown data

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. Conclusion

In this research, experimentation of statistical machine translation of English to Afaan Oromoo was conducted and a score of 17.74% was found. Although Afaan Oromoo is among resource-scarce languages (Kula et. al., 2008) of the world, the result of this experiment shows that the amount of data available can be used as a good starting point to build machine translation system from English to Afaan Oromoo.

Despite the fact that there is not any existing MT system for English to Afaan Oromoo language pairs with which one can compare the result of this system, the researcher believes that comparing the score with other existing systems' scores for other language pairs will enable one to judge the level of achievement. For instance, it has been reported in 2006 (http://www.statmt.org/jhuws/?n=Members.EvanHerbst) that the BLEU score for English to German and that of German to English was 17.76% and 25.54% respectively. Thus, one can conclude that this system performed not too low as compared to the systems built on sufficient amount of resource.

6.2. Recommendation

As the extension of the current work, the researcher recommends the following:

- While the currently reported scores are not state-of-the-art, the researcher is confident that further experimentation and the addition of more bilingual data will raise the accuracy level of this system. Therefore, the researcher strongly recommends the addition of more bilingual data for further experimentation.
- Only one reference translation is provided for evaluation of the system in this research. However, if more human references are used to calculate the BLEU

score, the scores will be higher (i.e., scoring one system with four human reference translations will increase the number of overlapping words versus a score calculated with one human reference translation). Therefore, evaluating the system using more reference translations is left for future work.

- The researcher noticed that the same Afaan Oromoo words in the corpus were considered by the system as different words due to spelling errors. Therefore, the researcher strongly recommends the development of spell checker for Afaan Oromoo that will help facilitate the document preparation.
- For a given language, accuracy of translating from it is different from translating into it. For example, while building translation systems for 11 languages, Koehn (2005) has proved that the score of translation from German is different from the score of translation into German. Thus, it is good to compare which direction of translation between English – Afaan Oromoo gives a better result using the given corpus.
- As the tools and techniques available for Indo-European languages are proved to be useful for Afaan Oromoo in this research, the researcher believes that these tools and techniques should be applied for other languages in Ethiopia to help the speakers of the languages reap the benefits of getting documents available in English without renouncing their own language.

Bibliography

Arnold, D., Lorna B., Siety M., Lee Sadler, R., Humphreys, L.. (1994). *Machine Translation: an Introductory Guide*. NCC Blackwell, London

Asefa W. (2005). *Development of Morphological Analyzer for Afaan Oromoo Text*. Unpublished MSc Thesis, AAU

Banerjee, S. & Lavie, A. (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*, Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics, Ann Arbor, Michigan.

Banerjee, S & Pedersen, T. (2003). *The Design, Implementation and Use of the Ngram Statistics Package*. In Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 03). Mexico City, Mexico.

Berger, A. (1997). *Trigger Toolkit*. Retrieved: Dec 25, 2008, From http://www-2.cs.cmu.edu/~aberger/software/TTK.html.

Berger, A. S, Pietra,S.A, Pietra, V.J. (1996). *A Maximum Entropy Approach to Natural Language Processing*. Retrieved: Dec 25, 2008. From: http://www.cs.cmu.edu/~aberger/maxent.html

Bernard E. Scott. (1998). *Linguistic and Computational Motivation for the LOGOS Machine Transition System*. LOGOS Co. Mt. Arlington, NJ.

Brown, P. F., J Cocke, S A Della Pietra, V J Della Pietra, F Jelinek, J D Lafferty, R L Mercer, and P S Roosin, P.S. (1990). *A statistical approach to machine translation*. Computational Linguistics, 16(2):79–85.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). *The mathematics of statistical machine translation: Parameter estimation*. ComputationalLinguistics, 19(2):263–311.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L., Cocke, J., Jelinek, F., Lai, J. (1995). *Method and system for natural language translation*. U.S. Patent 5,477,451.

Carl, M. and Sisay, F. (2003). *Phrase-based Evaluation of Word-to-Word Alignments. Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. HLT-NAACL Workshop.

Chaniak, E, Knight K, Yamada K. (2003). *Syntax-based Language Models for Statistical machine Translation*.

Chevalier, M.; Dansereau, J.; and Poulin, G. (1978). *TAUM-METEO: description du systme*. Groupe TAUM, Universit6 de Montrral, Montreal, Canada. Chevalier, Monique; Isabelle, Pierre; Labelle, Francois; and Lainr,

Clarkson, P.R., Rosenfeld, R.(1997). *CMU-Cambridge Statistical Language Modeling toolkit*. Retrieved: Dec, 23, 2008 From:

http://mi.eng.cam.ac.uk/~prc14/toolkit.html

Crego J. M., Costa-jussà M. R., Mariño J. B. and Fonollosa J. A. R. (2005). *Ngram-based versus Phrase-based Statistical Machine Translation*. TALP Research Center. Universitat Polit_ecnica de Catalunya, Barcelona

Croft, W. Bruce and Lafferty, John, ed. (2003). *Language Modeling for Information Retrieval*. Kluwer Academic Publishers.

Diriba M.. (2002). *An Automatic Sentence Parser for Oromoo Language*. Unpublished MSc Thesis, AAU.

Doddington, G. (2002). *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. Human Language Technology

Eisele, A, Federmann, C, Saint-Amand, H, Jellinghaus, M, Herrmann, T, Chen, Y. (2008). *Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System*. Preceedings of the Third Workshop on Statistical Machine Translation. Pp. 179-182. Columbus, Ohio, USA. ACL

Fishel, M, Kaalep H, Muischnek K. (2007). *Estonian-English SMT: the First Results*.

Germann, U. (2001). *Building a Statistical Machine Translation System from Scratch: How Much Bang for the Buck Can We Expect?*

Hiemstra, D. (1998). *Twente word alignment software*. University of Twente

Hersh, R. William. (2003). *Information Retrieval A Health and Biomedical Perspective*. Second Edition. Health Informatics Series.

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood, Chichester, England.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics, and Speech Recognition. Prentice-Hall.

Knight, K. (1999a). *A statistical MT tutorial workbook*. Unpublished. Retrieved August 27, 2008, from http://www.cisp.jhu.edu/ws99/projects/mt/wkbk.rtf.

Knight, K. (1999b). *Decoding complexity in word-replacement translation models*. Computational linguistics, 25(4), 607-615.

Koehn, P. (2003). *Advances in Statistical Machine Translation: Phrases, Noun Phrases and Beyond*.

Koehn, P. (2004). *Statistical significance tests for machine translation evaluation*. EMNLP-04, 388-395

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation*.

Koehn, P. (2006). *Statistical Machine Translation*, Draft Text Book. University of Edinburgh, Scotland.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). *Moses: Open source toolkit for statistical machine translation*. In Proc. of ACL Demo and Poster Sessions. 177–180.

Kula, K.,Varma, V. and Pingali, P. (2008). *Evaluation of Oromo-English Cross-Language Information Retrieval*, IIIT, Hyderabad, India.

Leon, M. (1984). *Development of English-Spanish Machine Translation*. Technical Report, Pan American Health Organization.Lippmann, E.O.

Locke, W.N and Booth A.D, (Editors). (1955). *Machine Translation of Languages*. New York, Wiley.

Lopez, A. (2008). *Statistical machine translation*. ACM Computing Surveys, 40(3)

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Morka M. (2001). *Text-to-Speech System for Afaan Oromoo*. Unpublished MSc Thesis, AAU.

Nogués, M. (2006). *Combining Machine Learning and Rule-Based Approaches in Spanish Syntactic Generation*

Och, F. J. & Ney, H. (2003). *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, 29(1):19–51.

Och, F J & Ney, H. (2004). *The Alignment Template Approach to SMT*. ACL

Papineni, K., S. Roukos, T. Ward, & W. Zhu. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. ACL

Pedersen, T & Varma, N. (2001). *K-vec++: Approach for Finding Word Correspondences*

Ramanathan, A. (2005). *Statistical Machine Translation* PHD Seminar Report. Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, Mumbai.

Rosenfeld, R, Chen,S.F., and Zhu, X. (2001). *Whole-Sentence Exponential Language Models: A Vehicle For Linguistic-Statistical Integration*. Retrieved: December 25, 2008. from: http://www.cs.cmu.edu/~roni/papers/wsme-csl-00.pdf

Roukos, S., Graff, D., and Melamed, D. (1997). *Hansard French/English*. Retrieved: Feb 20, 2009. From: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

Russell, Stuart and Norvig, Peter. (2003). *Artificial Intelligence A Modern Approach*. Second Edition. Pearson Education Inc.

Scannel, K. *Corpus building for minority languages*. Retrieved: August 01, 2008. From: http://borel.slu.edu/obair/om.html

Schultz, Tanja and Kirchhoff, Katrine. (2006). *Multilingual Speech Processing*. Academic Press, Elsevier Inc.

Sisay, F. and Haller, J. (2003a). *Amharic verbal lexicon in the Context of Machine Translation*. Traitement Automatique des Langues Naturelles TALN

Sisay, F. and Haller, J. (2003b). *Application of Corpus-based Techniques to Amharic Texts*, MT Summit IX Workshop Machine Translation for Semitic Languages: Issues and Approaches.

Slocum, J. (1984). *METAL: The LRC Machine Translation System*. Linguistics Research Center, University of Texas, Austin.

Smith, J. T. (2008). *Automatic Translation Vs Manual Translation*. Retrieved June 5, 2008 From: http://ezinearticles.com/?Automatic-Translation-Vs-Manual-Translation&id=1036194

Stolcke, A. (2002). *Srilm – an extensible language modeling toolkit*. In Proceedings of the International Conference on Spoken Language Processing, vol 2, pp 901-904. Denver, Colorado, USA.

Su, D. (2004). *Target-Dominant Chinese English*.

Toma, P. (1977). *SYSTRAN as a Multi-lingual Machine Translation System*.

Tilahun G. (1989). *Oromo – English Dictionary*

Tilahun G. (1993). *Qube Affan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet*. The Journal of Oromo Studies 1(1)

Turian, J., Shen, L., & Melamed, I. D. (2003). *Evaluation of machine translation and its evaluation*. In MT Summit IX.

Wakshum M. (2000). *Development of a Stemming Algorithm for Afaan Oromoo Text*. Unpublished MSc Thesis, AAU

Yihenew S. (2004). *Design and Development of Human-Aided Rule-Based English-Amharic Machine Translation Prototype*. Unpublished MSc Thesis, AAU

Zhang, M., Jiang H, Aw A. T., Sun J., Li S., Tan C. L. (2007). *A Tree-to-Tree Alignment-based Model for Statistical Machine Translation*.

Media. (2004). *Microsoft Encarta Encyclopedia*. Microsoft Corporation.

URL. (2008). *Machine Translation*. Retrieved June 5, 2008 From: http://en.wikipedia.org/wiki/Machine_translation/

URL. (2002). *Web Characterization Project*. Online Computer Library Center. Retrieved June 4, 2008, from http://www.oclc.org/research/projects/archive/wcp/stats/intnl.htm

URL. (2006). Open *Source Toolkit for Statistical Machine Translation*. Retrieved March 24, 2009 From: http://www.statmt.org/jhuws/?n=Members.EvanHerbst

URL. *Oromosoft*. Retrieved June 4, 2008, from http://www.oromosoft.com

URL. *Wikipedia Encyclopedia*. Retrieved August 15, 2008, from http://om.wikipedia.org

URL. *Department of Linguistics*, Addis Ababa University, Retrieved June 6, 2008, from http://www.aau.edu.et/faculties/linguistics%20Edited%20Page/thesis.htm

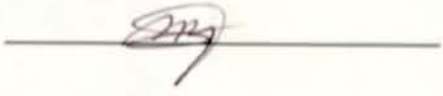
_____. (2002). *Statistical Abstract*. Central Statistical Authority, Federal Democratic Republic of Ethiopia

Appendix A: Sample output of the system

Sent	Systems	Content
1	Input	Proclamation No. 99/2005 Oromia National Regional Government Rural Land Use Payment and Agricultural Income Tax Amendment Proclamation.
	Reference	Labsii Lak. 99/1997 Labsii Kaffaltii Itti Fayyadama Lafa Baadiyyaafi Gibira Galii Hojii Qonnaa Mootummaa Naannoo Oromiyaa Irra Deebi'uun Fooyyessuuf Bahe.
	Output	Labsii Lak. 99/1997 Mootummaa Naannoo Oromiyaa Kaffaltii Itti Fayyadama Lafa Baadiyyaafi Gibira Galii Hojii Qonnaa ni karoorsa;.
2	Input	NOW, therefore, in accordance with article 49(3) (a) of the constitution of the Oromia National Regional Government it is hereby proclaimed as follows.
	Reference	Bu'uura Heera Mootummaa Naannoo Oromiyaa keewwata 49(3)(a) tiin kan kanatti aanu labsameera.
	Output	Akkaataa keewwata 49 (3) (a) tti kan Heera Fooyya'aa Mootummaa Naannoo Oromiyaa kan kanatti aanu labsameera.
3	Input	PART ONE General Provisions.
	Reference	KUTAA TOKKO Tumaalee Waliigalaa.
	Output	Kutaa Tokko Tumaalee Waliigalaa.
4	Input	2. Definitions In this proclamation unless the context requires otherwise.
	Reference	2. Hiikkaa Akkaataan seensa jechichaa hiikkaa biraa kan keennisiisuuf yoota'e malee labsii kana keessatti.
	Output	2. Hiika Labsii kana keessatti Hiika Akkaataan seensa jechichaa hiika biraa kan.
5	Input	But it does not include pastoralists.
	Reference	Haata'u malee horsiisee bulaa hin dabalatu.
	Output	Garuu daangaa hin dabalatu.
6	Input	3) Enterprise means state farm and any enterprise required to pay rural land use payment as per this proclamation.
	Reference	3) Dhaabbata jechuun Dhaaba Misooma Qonnaa qabiyyee Mootummaafi enterpiraayizii kamiyyuu kaffaltii itti fayyadama lafa baadiyyaa haala labsii kanaatiin kaffaluun irra jiru jechuudha.
	Output	(3) fi (dhiqamuu'jechaa Mootummaarraa kamiyyuu akka galiifi Kaffaltii itti fayyadama lafa baadiyyaa akkaataa Labsii kana.
7	Input	22. Powers and Duties of Trade, Transport and Industry Bureau.
	Reference	22. Aangoofi Hojii Biiroo Daldala, Geejjibaafi Industirii.
	Output	22. Aangoofi Hojii Biiroo Daldala,.
8	Input	2) To ensure the observance of laws and directives issued to regulate transport and trade services, supervise the trade and transport activities not to be implemented illegal.
	Reference	2) Seeronni, danboonniifi qajeelfamoonni hojii daldalaafi geejjibaa to'achuuf bahan hojiirraa ooluusaanii ni mirkaneessa, hojiin daldalaafi tajaajila geejjibaa seeraan ala akka hin rawwatamne ni to'ata.
	Output	2) Seeronni bahan akka hin rawwatamne fi daldalaafi geejjiba,; hojii daldalaafi geejjibaa akka hin buusu.
9	Input	The Bible urges us to show love and to be forgiving.—(Colossians3:12-14).
	Reference	Macaafni qulqulluun jaalala akka agarsisnuufi akka waliif dhiifnu cimsee nugorsa.—(Qolosaayis3:12-14).
	Output	Macaafni Qulqulluun wangeelli jaalala nuagarsiisuu fi forgiving â "(Qolosaayis3:12-14).
10	Input	Each year in the United States, almost 180,000 women are diagnosed with breast cancer.
	Reference	Baruma baraan biyya Ameerikaa kessatti dubartota 180,000 naqarsaa harmaa (golfee) godhatee muldhata
	Output	Waggaa tokko keessatti, tokkummaa qabaata namoonii 180,000 diagnosed breast waajjin kan dhorkamani dha.

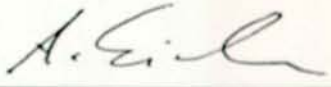
DECLARATION

The thesis is my original work, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.



Sisay Adugna Chala

The thesis has been submitted for examination with our approval as university advisors.



Dr. Andreas Eisele, Universität des Saarlandes/ DFKI GmbH, Germany

Ato Ermias Abebe, Addis Ababa University, Ethiopia