Addis Ababa University

Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

English-Amharic Document Translation

Using Hybrid Approach

By

Samrawit Zewgneh

A Thesis Submitted to the School of Graduate Studies

in Partial Fulfillment of

Masters of Science in Computer Engineering

Thesis Advisor Dr. Surafel Lemma

Chairman of Department

Dr. Yalemzewud Negash

Thesis Evaluator

Thesis Evaluator

Thesis Evaluator

English-Amharic Document Translation Using Hybrid Approach

By

Samrawit Zewgneh

MSC Thesis

Submitted in Partial Fulfillment of the Requirements

for Masters of Science in Computer Engineering

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

October, 2017

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with proper citation of sources. I, the undersigned, declare that this thesis has not been presented for a degree in any other university.

Samrawit Zewgneh

Acknowledgement

First of all , I would like to thank God for blessing my health and make me courageous during my study. I am grateful to my advisor Dr.Surafel Lemma, without his timely and valuable comments, advice and reassurance, this work would not be complete. I also indebted to acknowledge all my friends who involved in providing information used in this project. Finally , Jigjiga university and Addis Ababa Institute of Technology should be acknowledge for sponsoring and financial support respectively.

Abstract

Machine Translation (MT) is a sub field of Natural Language Processing (NLP) that investigates the use of computer software to translate text or speech from one natural language to another. Recently, in Ethiopia English documents are translated to Amharic using human translators, because of the absence of Automated Translation System. Due to this, the process of document translation is so expensive, challenging, unsecured and time consuming. To solve those problems statistical based approaches were proposed to translate English to Amharic. However, the approaches have accuracy and understandability issues. In order to solve these problems we propose a Hybrid approach machine translation (HMT) system that combines Statistical and Rule Based Machine Translation approaches.

In this case using the hybrid approach achieved better accuracy and found to be advanced for English - Amharic machine translation system over SMT approach. The proposed hybrid approach achieved 15% and 20% accuracy improvement for simple and complex sentences over statistical machine translation approach. This study also identified preprocessing the inputs of SMT is more suitable to improve accuracy for complex sentences while post-processing the outputs of SMT is more suitable for simple sentences.

Key Words: *Machine Translation, SMT, RBMT, HMT.*

Contents

List of Figures

List of Tables

Acronyms

BLEU	Bilingual Evaluation Understudy
DC	Dependent Clause
HMT	Hybrid Machine Translation
IC	Independent Clause
NLP	Natural language Processing
MLE	Maximum Likelihood Estimate
MT	Machine Translation
PBT	Phrase Based Translation
RBMT	Rule Based Machine Translation
SMT	Statistical Machine Translation
SOV	Subject Object Verb
SVO	Subject Verb Object
SWBT	Single Word Based Translation
POS	Part of Speech
QW	Question Word

Chapter 1

Introduction

Since the age of human being, languages had been processed and documented in various ways. Different literatures such as fiction, poetry, drama, criticism, scientific and cultural issues could be an evidence of documents written in several natural documents. To address those literatures to different language speakers, people have traditionally used a human document translator. However, this kind of approach is challenging, expensive and time consuming (Hutchins, 1995). To address these problems, experts have designed an automated Machine Translation system (Hutchins, 1995).

Machine Translation (MT) is a sub field of Natural Language Processing (NLP) that investigates the use of computer software to translate text or speech from one natural language to another (Teshome, 2013). Machine translation is designed to speed up the rate of document can be translated, as well as to reduce overall costs of translation process (Slocum, 1984). Various methodologies have been devised to automate the translation process to restore the meaning of the original text in to translated version. The most known and applicable methods of translation system are *Statistical*, *Rule Based* and *Hybrid* approaches (Costa-Jussa et al., 2012).

Statistical Machine Translation (SMT) deals with automatically mapping sentences in one human language (e.g, English) into another human language (e.g, Amharic). The first language is called the source and the second language is called the target. This process can be thought as a stochastic process (Costa-Jussa et al., 2012). There are many SMT variants, depending upon how translation is modeled. Some variants are based on string-to-string mappings while others are based on tree-to-string or tree-to-tree mappings (Osborne, 2011).

Statistical Machine Translation has proven success in the linguistic technology. State of

the art translation tools like Google Translate and Bing Translator are built based on statistical machine translation (España Bonet et al., 2011). The scope of the translation ranges from word level, to sentence and even to document level. Even though most of the work so far has focused on sentence level translation (Osborne, 2011).

Rule-based machine translation (RBMT) systems were the first commercial machine translation systems (Osborne, 2011). RBMT is much more complex than translating word to word. These system develop linguistic rules that allow the words to be put in different places, to have different meaning depending on context (Osborne, 2011). RBMT systems are built on large size dictionaries and sophisticated linguistic rules. Users can improve translation quality by adding terminology into the translation process (Thurmair, 2004). In general RBMT systems produce syntactically and grammatically better translations and deal with long distance dependencies, agreement and constituent reordering in a better way (España Bonet et al., 2011).

Hybrid Machine Translation Approach(HBMT), takes the advantage of both statistical and rule-based translation methodologies. This approach has proven better efficiency in the area of MT systems (Okpor, 2014). Hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information which is called Guided by RBMT (Okpor, 2014). In the other cases, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system which is called Guided by SMT. This technique is better than the previous and has more power, flexibility, and control in translation (Okpor, 2014).

In Ethiopia documents are translated using Human translators because of the scarcity of Automated Translation System. Due to this, the process of document translation is expensive, challenging, unsecured and time consuming. To alleviate such problems,

Ambaye and Eleni (Tadesse, 2012) (Teshome, 2013) proposed Amharic-English Machine translation system using Statistical Machine Translation approach. However, the proposed approaches, have problems on accuracy and understandability. Understandability tells us is the translated sentence is understandable or not. The researches are done by using only SMT approach, but we believe that the accuracy of English-Amharic MT could be further improved by applying advanced machine translation approach such us *Hybrid MT approach and Machine Translation using Deep Neural Network*. Therefore this study is planned to address those problems and to develop English-Amharic translation system by using hybrid approach.

1.1 Statement of the problem

Many literatures show that there are problems in English-Amharic Machine Translation in terms of understandability and accuracy (Clemente, 2008). Those problems arise because of lexical complexity and shortage of resources that guide the translation system in Amharic language. Ambaye and Eleni have done an experiment on English-Amharic Machine translation system based on SMT (Teshome, 2013). However, their works, had problems on accuracy and understandability for complex sentences. The approach used in their experiment, SMT also has its own weakness on selection of grammatical and syntactical structure for complex sentences as well as simple sentence to some extent (España Bonet et al., 2011). Therefore, the main aim of this study is to reduce the complexity of SMT approach using hybrid approach which combines SMT and RBMT. We also investigate if the hybrid approach is better on selection of grammatical and syntactical structure for complex sentences than SMT for English-Amharic MT system? In particular, we investigates the following research questions (RQ)s:

RQ1: Is hybrid MT approach better than SMT approach for English-Amharic MT system ?

RQ2: Which RBMT component has better contribution to the hybrid system ?

1.2 Objective

General Objective

The main objective of this study is to design and develop English-Amharic document translation system by using hybrid approach.

Specific Objective

- To implement SMT based Machine translation for English-Amharic.
- To design and implement hybrid(SMT with RBMT) based machine translation for English-Amharic.
- To compare the result of both systems and conclude.

1.3 Significance of the Study

MT systems play great role for one language beside minimization of cost and time consumption. This study has the following benefits.

- To facilitate technology transfer from developed countries.
- To increase the acceptability of new finding and technologies.
- To decrease the cost and time of document translation process.
- To convert technical documentation which are written in English language into Amharic language.
- For maintaining confidentiality and security during translation of documents.

1.4 Research methodology

Several different methodologies have been used to hybridize MT approaches. Hybridization of RBMT and SMT can be classified into guided by RBMT and guided by SMT

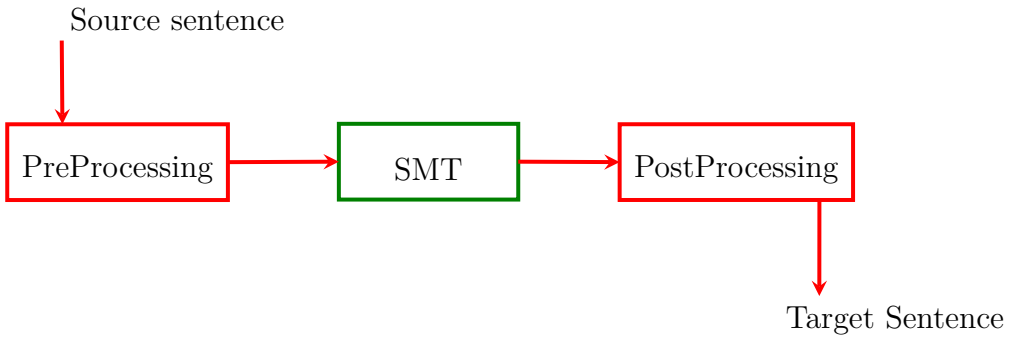

Figure 1.1: Hybrid Approach Architecture

(Costa-Jussa and Fonollosa, 2015). The former integrates data information into a rule-based architecture; the latter integrates linguistic rules into SMT architecture (Costa-Jussa and Fonollosa, 2015). On this study, guided by SMT hybridization is used as shown in Figure 1.1. The first step is Pre-processing, that used to reorder the source sentence into a form that better matches to the target sentence. Then the SMT system takes reordered sentence as an input and produces an output. Finally Post-processing is used for checking the output of SMT. If the output sentence is grammatically correct the target sentence will be delivered, otherwise the grammar corrector corrects the grammar and deliver the corrected sentence. The source and target sentences for this study are English and Amharic languages respectively (Teshome, 2013).

In general the overall methodology of this study is as follows:

- Phase 1: Getting better understanding about both languages.
- Phase 2: Collect electronic text corpus in English-Amharic
- Phase 3: Collect Grammar rules for Amharic and English languages.
- Phase 4: Implement English-Amharic MT system using SMT.
- Phase 5: Design and implement English-Amharic MT system using our Hybrid approach.

- Phase 6: Test both systems independently.
- Phase 7: Compare the result of both systems.

1.5 Scope

This study deals with designing and implementing automated English-Amharic Machine Translation using hybrid approach. Even though there is a bidirectional translation system, the current study is limited on unidirectional system which is English to Amharic translation.

1.6 Thesis outline

Chapter 2 provides background information about machine translation approaches in detail. In *Chapter 3* we discuss the work of related researches in the area of English-Amharic machine translation and other languages. The proposed hybrid approach is discuss in *Chapter 4*. The evaluation of proposed approach is discuss in *Chapter 5*. The result and discussion of the evaluation are presented in *Chapter 6*. Finally conclusion and recommendation is provided in *Chapter 7*.

Chapter 2

Background

In this chapter, we are going to discuss background information about Machine translation and its approach in detail. In addition we discuss about strength and weakness of SMT and RBMT approaches.

2.1 Machine Translation

Machine translation is a sub field of Natural Language Processing(NLP) that studies the use of computer to translate text of speech from one target language to another (Teshome, 2013). Machine translation is also a subfield under Artificial Intelligence (Okpor, 2014).

Recently, there are many approaches used in machine translation world. The most known approaches are Statistical (corpus based), Rule based (knowledge based) and Hybrid approaches. We will see those approaches one by one in the coming sections.

2.1.1 Statistical

In 1949, Warren Weaver had come up with the idea of Statistical Machine Translation (SMT) (Tripathi, 2010). The statistical methods are applied to perform translation by using models estimated from parallel corpora (Tripathi, 2010). The process of translating one language from another language is called stochastic process (Osborne, 2011). SMT can be classified depending on their translation models such as *Statistical Word Based Translation Model, Statistical Phrase Based Translation Model and Statistical Syntax Based Translation Model* (Tripathi, 2010).

Officially, SMT translation can be thought as searching the most acceptable target sentence A* for source sentence E :

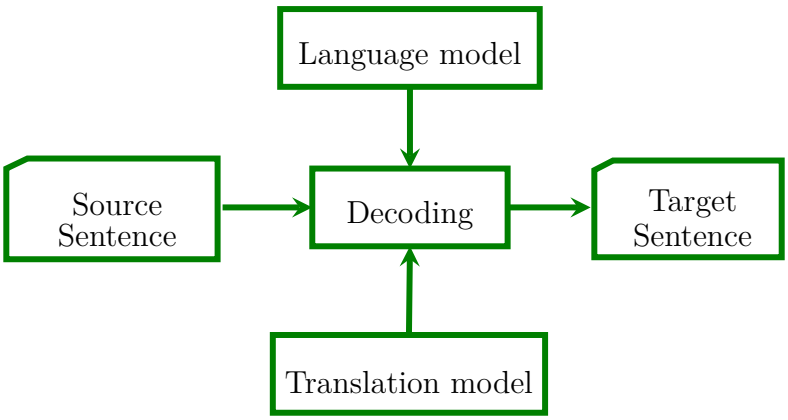

Figure 2.1: Architecture of SMT

$$A^* = argmax P(E|A)P(A) \tag{2.1}$$

Where, A stands for Amharic and E for English, but any language pairs can be substituted. $P(A)$ is the *language model*, and $P(E|A)$ is the *translation model*. The language model is used for assigning the probability $P(A)$ to each Amharic sentence A (Knight, 1999). The translation model assigns a probability that a given source language sentence E generates target language sentence A. Which means, modeling the transformation probability from E sentence to A sentence. We can see the architecture of SMT in Figure 2.1

2.1.2 Rule Based

Rule based approach machine translation system is a system that is depending on linguistic rules and knowledges of source and target languages. The linguistic information about the source and target languages are retrieved from bilingual dictionaries and grammars covering of main semantic, morphological, and syntactic regularities (Okpor, 2014).
RBMT needs a high variety of linguistic phenomena. Grammar exceptions will have big problems on RBMT systems (Tripathi, 2010). Having input sentence in some source language (for example, English) RBMT system will generate or translate the sentence to some target language (for example, Amharic) on the basis of syntactic, semantic and morphological analysis of both source and target languages.

2.1.3 Hybrid Approach

Hybrid approach integrates both statistical and rule-based MT methodologies. The approach takes advantage and weakness of both approaches. Hybrid Machine Translation (HMT), has proven better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use hybrid based approach to develop translation from source to target language (Okpor, 2014).

There are some methodologies which have been used to hybridize (integrate) different MT within and across paradigms. As shown in Figure 2.2, hybridization of RBMT and SMT can be classified into **Guided by RBMT** and **Guided by SMT**. Guided by RBMT incorporate data information into a rule-based architecture. Guided by SMT incorporate linguistic informations into SMT architecture (Costa-Jussa and Fonollosa, 2015).

Guided by RBMT

In this hybridization method the main system is developed by Rule based approach, then SMT approach is added on pre/post processing stage to improve the efficiency of RBMT output. There are three kinds of strategies within this category:

1. Introducing corpus to build the RBMT system
2. Introducing SMT tools to weight the RBMT output
3. Carrying out a statistical post-editing of a RBMT output

The first strategy, *introducing corpus to build the RBMT system*, i.e., generating lexical, syntactical and morphological rules from the parallel corpora. The main reason for using data when building RBMT system is to reduce its cost, time and effort required for building RBMT system (Costa-Jussa and Fonollosa, 2015).

The second strategy is, *introducing SMT tools to weight the RBMT output.* This strategy is focuses on the RBMT output by integrating tools such as language models or stochastic

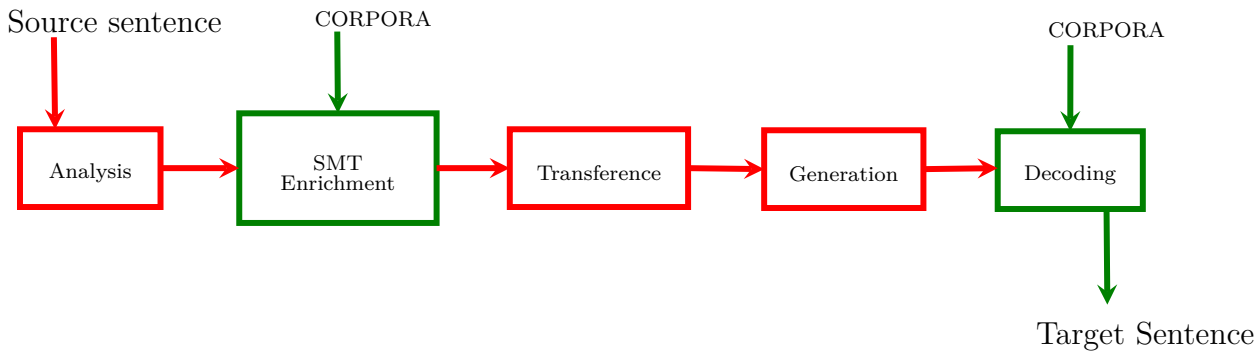

Figure 2.2: Architecture of Hybrid Approach guided by RBMT

parsers to provide better translated output (Costa-Jussa and Fonollosa, 2015).

The third strategy is *carrying out a statistical post-editing of RBMT output*. Which considers RBMT outputs as source sentences and post-edited results by using SMT approach as target sentences. Only in this strategy, RBMT and SMT paradigms are concatenated but not integrated at the architecture level (Costa-Jussa and Fonollosa, 2015).

Guided by SMT

The other hybridization method is guided by SMT. In this case the main system is developed by SMT approach, then RBMT system is integrated to improve the efficiency of SMT output as shown in Figure 2.3 . The integration could be done in two ways:

1. Using rules at pre/post-processing

2. Integrating dictionaries/rules into the core model of SMT

The first technique is *adding rules in pre/post processing* stages. Adding rules in preprocessing stage means, syntactical or semantical rules have been used to reorder the source sentence into a form that better matches for target sentence. On the other hand adding rules in post processing means, rules have been used for morphology generation by means of combining machine learning and introducing dictionaries (Costa-Jussa and Fonollosa,

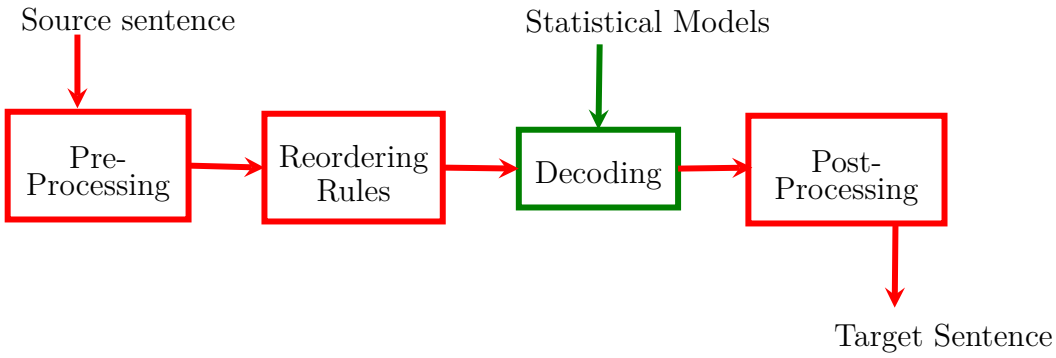

Figure 2.3: Architecture of Hybrid Approach guided by SMT

2015).

The second technique is *integrating dictionaries/rules into the core model of SMT*. This kind of technique integrates morphological or syntactical rule to the core model of SMT system. Using RBMT information to improve statistical word alignment can be an example.

In general, hybridization methods should be chosen depending on the need of MT system and on the nature of the source and target languages.

2.1.4 Comparison of RBMT and SMT systems

SMT

- Better for user generated content and broad domain materials.
- Will use the most likely term but not necessarily the one you wanted.
- The system is more robust and always produce output.
- Have difficulties to cope with phenomena that requires linguistic knowledge.
- Requires significant hardware to build.
- SMT system is cost effective than other MT systems.

RBMT

- Translate more accurately by trying to represent every pieces of the input.
- Have predictable output.
- Better suited to post editing and durable change.
- Expensive to build the system , because it needs high linguistic knowledge about the source and target languages.

Chapter 3

Literature review

Machine translation (MT) is a process of translating one language to another without the interference of human being. Machine translation can be implemented using three different approaches as discussed in previous Chapter. We classified the related literatures based on the three approaches: SMT, RBMT and Hybrid based.

3.1 SMT Approach

There are many researches that are done on Amharic language processing. However, only three studies are done on English-Amharic document translation. Ambaye and Yared (Tadesse, 2012) developed the first Amharic-English Machine translation system. They used Statistical Machine translation approach. They got 28.4% and 11.43% accuracy for phrase based and hierarchical based translation respectively. The result shows that there is high accuracy problem (Tadesse, 2012).

To solve Ambaye and Yared (Tadesse, 2012) accuracy problem, Eleni (Teshome, 2013) had done an experiment on Amharic-English translation system using Constrained Corpus which have better accuracy (Tadesse, 2012). She used two different corpus; for simple and complex sentences. The result she obtained was an average of 82.22% and 73.38% accuracy for simple and complex sentence respectively. As the result shows, she got better accuracy compared to Ambaye and Yared(Tadesse, 2012) work. We believe that this accuracy could be further improved by applying advanced machine translation approach.

Another experiment were done on the same year by Teshome (Teshome and Besacier, 2012). He had conducted a preliminary experiment on translation of English to Amharic using the Statistical Machine Translation (SMT) approach. The experiment has been conducted using 18,432 English-Amharic sentence pairs. The baseline phrase-based BLEU

score result is 35.32%. A 0.34% increase in BLEU has been achieved by applying morpheme segmentation to the tokens of the Amharic output result and the reference of the baseline system. The increase is 0.92% when compared with the same segmented reference between the baseline and the segmented system.

In general, the above three studies used statistical machine translation approach. But SMT approach has weakness on selection of grammatical and syntactical structure for complex sentence as well as for simple sentence (España Bonet et al., 2011). Therefore, in this research, we plan to investigate if a hybrid approach that combines statistical and Rule Based Machine Translation approach would address the aforementioned problems.

3.2 RBMT Approach

Rule based MT system for English-Amharic seems hard to develop as (Teshome et,al) (Teshome and Besacier, 2012) mentioned, because of the scarcity of documented linguistic resource in Amharic. Despite the difficulty, Gasser (Gasser, 2012) developed a rule based English-Amharic MT using L3 frameworks. The development is still on going. He already introduced an L3 framework for the development of rule based English-Amharic MT system (Gasser, 2012).

3.3 Hybrid Approach

Several researchs use hybrid approach to translate English to Ethiopian languages. One of them is *"Bidirectional English Afaan Oromo Machine Translation Using Hybrid Approach"* (Daba, 2013). This study resulted in the development of a bidirectional English-Afaan Oromo machine translation system using a hybrid approach. Jabesa Daba (Daba, 2013) had done two experiments in both SMT and Hybrid approach.

The first experiment is carried out by using a **statistical** approach. The result obtained from the experiment has a BLEU score of 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation.

The second experiment is carried out by using a **hybrid** approach and the result obtained from the experiment was BLEU score of **37.41%** for English to Afaan Oromo translation and **52.02%** for Afaan Oromo to English translation.

From the result, we can see that the hybrid approach is better than the statistical approach for the language pair and improved translation is acquired when Afaan Oromo is used as a source language. In this literature the rule based part is added only on preprocessing stage, i.e., only on inputs of SMT system. The system uses only syntax reordering to pre-process inputs of SMT.

The above "Bidirectional English Afaan Oromo Machine Translation Using Hybrid Approach" research has some similarities with the proposed English-Amharic hybrid approach on preprocessing stage. However, the rules used to pre-process the inputs of SMT is different. The current proposed English-Amharic MT system using hybrid approach also has additional *post processing stage* that applied for correcting the outputs of SMT to improve the translation process. We also developed languages rules that are specific to Amharic.

Chapter 4

Hybrid approach

This Chapter discusses the over all architecture and components of the proposed hybrid approach for English-Amharic system. The overall architecture which is designed for English Amharic machine translation using hybrid approach is shown in Figure 4.1.
The input and output sentence in this study are English and Amharic sentences respectively. The broken line in the architecture shows RBMT while the normal line shows SMT. The proposed hybrid system has three main stages *pre-processing, decoding and post processing*. The approach is based on the concept of guided by SMT hybridization method as discuss in Chapter two. The main system is developed by SMT approach, but there are two additional Rule based components before the input of SMT and after the output of SMT system to improve the translation quality of SMT output.

4.1 Pre-processing

Preprocessing is important step in machine translation process, where source words reordering takes place based on linguistic information. This process aims at reducing the complexity of the translation process derived from structural differences (word order) of language pairs. One of the main problem in SMT approach is reordering source sentence structure to target sentence structure (Clemente, 2008). In this study two essential steps *Part of Speech Tagging* and *Structural Recording* have been followed to pre process the inputs of SMT.

4.1.1 Part of Speech Tagging

Tagging is a process of adding part-of-speech for each word in a sentence. Parts-of-speech are useful because of the large amount of information they give about a word

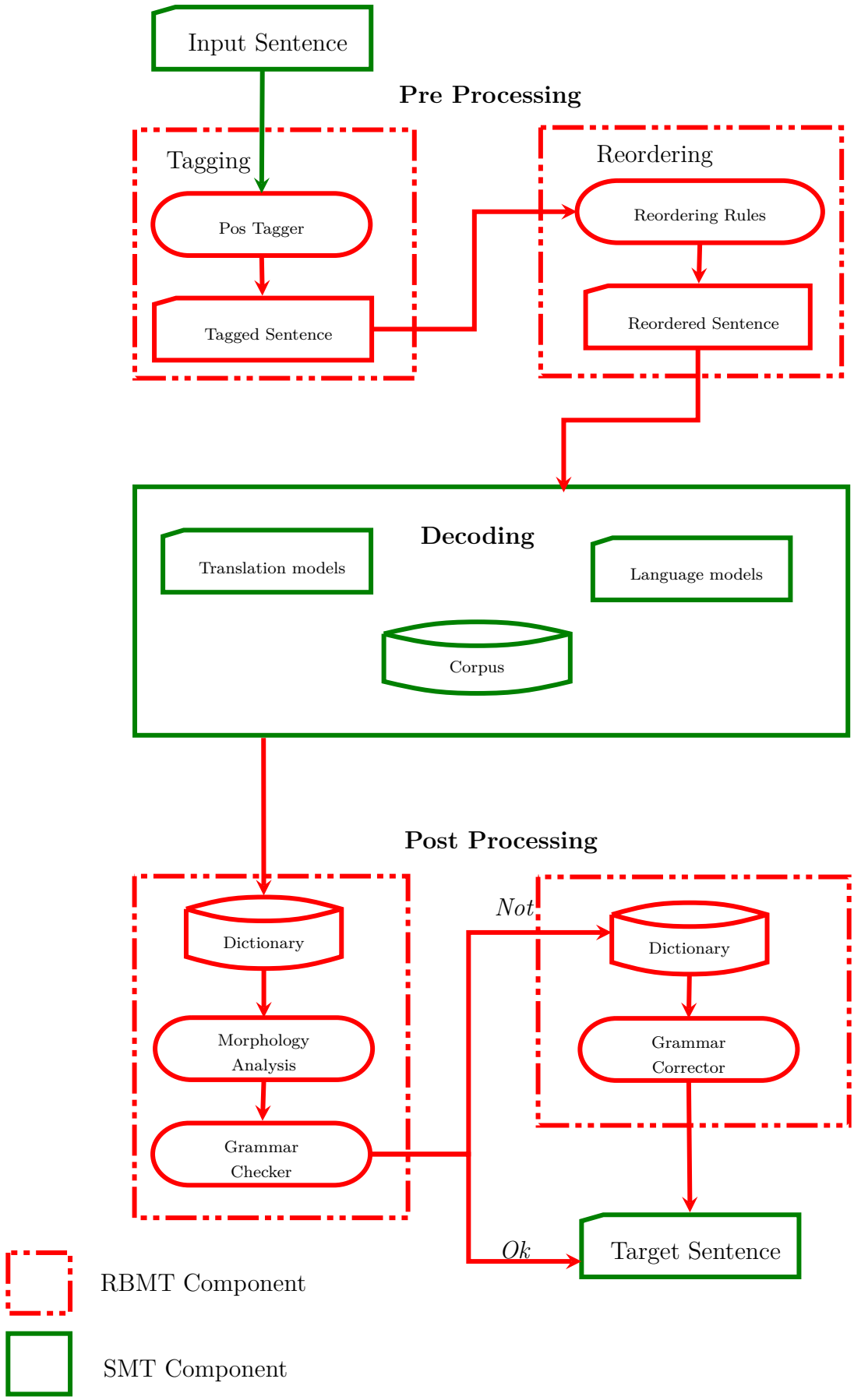

Figure 4.1: Hybrid System Architecture

Tag	Description	Example	Tag	Description	Example
CC	Coordin, conjunction	and, but, or	RP	Particle	Up, of
CD	Cardinal number	one, two	SYM	Symbol	+, %, &
DT	Determiner	a, the	TO	"to"	to
EX	Existential "there"	there	UH	Interjection	Uh, Oops
FW	Foreign word	fora cupa	VB	Verb base form	eat
IN	Preposition /sub-conj	of, in, by	VB	Verb past tense	ate
JJ	Adjective	Yellow	VBG	Verb gerund	eating
JJR	Adjective comparative	bigger	VBN	Verb past participle	eaten
JJS	Adjective superlative	wildest	VBP	Verb non 3sg pres	eat
LS	List item marker	1,2	VBZ	Verb 3sg pres	eats
MD	Modal, can	should	WDT	WH determiner	which, that
NN	Noun sing, or mass	lemma	WP	Wh-pronoun	what, who
NNS	Noun plural	lemmas	WPS	Possessive Wh	whose
NNP	Proper noun, sing	IBM	WRB	Wh-adverb	where, how
NNPS	Proper noun, plural	Carolinas	$	Dollar sign	$
PDT	Pre-determiner	all, both	#	Pound sign	#
POS	Possessive ending	's	"	Left quote	"
PRP	Personal pronoun	I, you, he	"	Right quote	"
PRP$	Possessive pronoun	your	(	Left parenthesis	[, (, f
RB	Adverb	quickly, never	)	Right parenthesis	],), g
RBR	Adverb comparative	faster	,	Comma	,
RBS	Adverb superlative	fastest	.	Sentence final punc	. ? !

Table 4.1: Penn Treebank part-of-speech tags (including punctuation)

and its neighbors. Knowing whether a word is noun or verb tells us a lot about likely neighboring words.

Tagging is the first step in pre processing which assigns parts-of-speech to each word for source sentences (English). On this study Stanford POS Tagger is used (Jurafsky and Martin, 2014). Stanford POS Tagger is freely available POS tagger which is available for English. While there are many lists of parts-of-speech, most modern language processing on English uses the 45-tags of **Penn Treebank** tag set (see Table 4.1) (Jurafsky and Martin, 2014). For example, if the input sentence is *"Which one is your brother"* then the Stanford POS tagger will assign part of speech to each word in the sentence and generate the given sentence, *"Which_WDT one_NN is_VBZ your_PRP$ brother_NN"*.

4.1.2 Structural Reordering

Reordering is one of important steps in Hybrid approach, where a source sentence structure (word order) is reordered as the target sentence structure. This process aims at reducing the complexity of the translation process derived from structural differences (word order) of language pairs (Clemente, 2008). Because SMT system generates its reordering table during the training stage from the parallel corpus. The main objective of reordering is reducing the reordering errors which appear in decoding process of SMT by applying linguistic informations (Reordering rules) of both source and target languages.

The major structural difference between English and Amharic language is that English follows **Subject-Verb-Object (SVO)** sentence structure, whereas, Amharic language follow **Subject-Object-Verb (SOV)** sentence structure. Because of the difference in word order, words in the source sentence are aligned to target sentence words which have a different position. The above aforementioned reason leads us reordering is needed for English to Amharic MT system.
We have collected some reordering rules for both Amharic and English languages from different grammar books (Kim and Sells, 2008) (Piscioneri et al., 2011) (Getahun, 2010) (Baye, 2000). Those rules are applied to reorder English sentence to the structure of Amharic sentence. The reordering process helps the decoding process to search best match of the target language (Amharic) from the trained corpus. The reordering rules are classified in to three category.

A Reordering rule for simple sentence

B Reordering rule for question sentence

C Reordering rule for complex sentence

A. Reordering rule for simple sentence

1. Verb Reordering Rule

Verbs are one of English word classes that describe an action of something. In a declarative sentence, the subject of a sentence comes directly in front of a verb. The direct object comes directly after the verb. On the other hand, in Amharic Language the subject of a sentence comes directly in front of the direct object (when there is one). The verb comes directly after the direct object.

Example 4.1

The man wrote a letter = ሰውየው ደብዳቤ ጻፈ

"The man" is a subject "wrote" is a verb and "a letter" is direct object of English sentence. "ሰውየው" is a subject, "ደብዳቤ" is direct object and "ጻፈ" is a verb for Amharic sentence. From this we can see the structure of the two sentences are different.

Hence, we are going to reorder the English sentence to the structure of Amharic sentence as follows by using **Algorithm 1**.

The man a letter wrote. = ሰውየው ደብዳቤ ጻፈ

Algorithm 1 Reordering verbs in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        if VB in w_j then
            (tmp) = w_j
            w_j = w(length (s)-1)
            w(length(s)-1) = (tmp)
            store s_i in S
        end if
    end for
end for
Write S to file
```

Preposition in Amharic language	**Description in English**
Le	to , for
Be	at, in, on, with, by
Ke	from, at, than
Wede	to, towards
Sele	Becouse, about, forthe sake of concering
Ende	as, like, according to
Yale	without
Ye	of,....'s

Table 4.2: Preposition of Amharic language

Preposition	**Description**
Prepositions of time:	after, around, at, before, between, during, from, on, until, at, in, from, since, for, during, within
Prepositions of place:	above, across, against, along, among, around, at, behind, below, beneath, beside, between, beyond, by, down, in, inside, into, near, off, on, opposite, out, over, past, through, to, toward, under, underneath
Prepositions of direction/movement:	at, for, on, to, in, into, onto, between
Prepositions of manner:	by, on, in, like, with
Other types of prepositions:	by, with , of, for, by, like, as

Table 4.3: Preposition of English language

2. Preposition phrase Reordering Rule

Prepositions are words that show the relationship between a noun or a pronoun and some other words or elements in the sentence. A phrase is a group of words working together that does not have both a subject and a verb. Phrases usually act as a single part of speech. A prepositional phrase is a phrase that begins with a preposition and ends with noun, pronoun and gerund. In Amharic language prepositions (መስተዋድድ) are usually inseparable from the word they precede. Tables 4.2 and 4.3 show prepositions of Amharic and English languages respectively.

Position of a preposition in prepositional phrase is before noun, pronoun and a gerund. In Amharic language, some prepositions come before noun, pronoun and gerund, while

other prepositions comes after the noun, pronoun and gerund.

Some prepositions that come after a noun and pronoun in Amharic language are *after, around, at, before, between, during, until, within, above, against, below, beyond, down, inside, out, over, toward, underneath.*

Example: 4.2

After everybody => ከሁሉም በኋላ

Example 4.3

Before him. => ከሱ በፊት

The above English phrases and their translations in Amharic shows, the prepositions (underline words) appear before the nouns and pronouns in English. In Amharic, however, the prepositions appear after the pronouns. Therefore, in order to avoid such structural difference we have reordered English phrase by using **Algorithm 2**. After applying the reordering rules to the above two phrases we can get the following reordered phrases:

Everybody after => ከሁሉም በኋላ

Him before => ከሱ በፊት

3. Preposition with cardinal number reordering rule

A cardinal number is a number such as 1, 3, or 10 that tells how many things there are in a group but not their order. In English sentence structure prepositions are placed before cardinal numbers. On the other hand, in the Amharic sentence structure cardinal numbers come before preposition. But those only those prepositions listed in Section 2.

Algorithm 2 Reordering prepositional phrases in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
Store (after, around , at, before, between, during,
until, within, above, against, below, beyond, down, inside, out,
over, toward, underneath) in P
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        for each p_n ∈ {P} where n=0,1,2,3...n do
            if VB and p_n in w_j and (NNS or NNSP or NNP or NN) in w_{j+1} then
                tmp = w_j
                w_j = w_{j+1}
                w_{j+1} = tmp
                store s_i in S
            end if
        end for
    end for
end for
Write S to file
```

Example 4.4

Before two => ከሁለት በፊት

In the first English phrase preposition ”before” appear before cardinal number ”two”. Whereas in the Amharic phrase preposition “በፊት” appear after cardinal number “ከሁለት”. Hence, in order to avoid such structural (word order) difference we have reordered the English phrase to the Amharic phrase structure. **Algorithm 3** is used to reorder this kind of phrases. After applying the reordering rule the above two phrases will be as follows:

Two before. => ከሁለት በፊት

Algorithm 3 Reordering preposition with cardinal number in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
store (after, around , at, before, between, during,
until, within, above, against, below, beyond, down, inside, out,
over, toward, underneath) in P
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        for each p_n ∈ {P} where n=0,1,2,3...n do
            if IN and p_n in w_j and CD in w_{j+1} then
                tmp = w_j
                w_j = w_{j+1}
                w_{j+1} = tmp
                store s_i in S
            end if
        end for
    end for
end for
Write S to file
```

4. Adjectives with verb Reordering Rule

An adjective is a word that describes a noun. Adjectives describe nouns by giving some information about an object's size, shape, age, color, origin or material. In English sentence structure adjectives usually come directly before noun and predicate position and also directly after a verb, but this is true for few adjectives like: *aware, alive, asleep and awake* that can only be used after a verb. Whereas in Amharic sentence structure (word order) those adjectives are followed by verbs.

Example 4.5

Becoming difficult => ከባድ እየሆነ

In the first English sentence the verb "Becoming" appear before adjective "difficult". Whereas in the second Amharic sentence the verb "እየሆነ" appear after adjective "ከባድ". Therefore, in order to avoid such structures (word order) difference we have reordered the English sentence to the same as the Amharic sentence structure. **Algorithm 4** is used for reordering this type of sentences. After applying the reordering rule the above two sentences will be as follows:

difficult Becoming => ከባድ እየሆነ

Algorithm 4 Algorithm for adjective verb reordering in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        if (VB or VBG or VBZ or VBN or VBP) in w_j and JJ in w_{j+1} then
            tmp = w_j
            w_j = w_{j+1}
            w_{j+1} = tmp
            store s_i in S
        end if
    end for
end for
Write S to file
```

5. Reordering rules for command

Command, which can be a noun or a verb, combines the Latin prefix *"com"*, meaning "with", and *"mandare"* , meaning "to charge or enjoin" giving someone a command is to say something with the authority that charges him to follow it. This kind of sentence in English follows a different word order with the declarative sentence structure which is (Subject-Verb-Object). In command sentence verbs usually come before the subject. On the other hand, in the Amharic language sentence structure, the position of a verb is always after the subject whether the sentence is command or not. But there is intonation change when we use command in Amharic

Example 4.6

Open the door ! => በሩን ክፈተው !

The above Example 4.6, in the first English sentence the verb "Open" appear before the subject "the door". Whereas in the second Amharic sentence the verb "ክፈተው" appear after the subject "በሩን". Therefore, in order to avoid such word order difference we have reordered the English sentence to the same as Amharic sentence structure. **Algorithm 5** is used for reordering this type of sentences. After applying the reordering rule the above two sentences will be as follows:

the door Open ! => በሩን ክፈተው !

Algorithm 5 Algorithm for Command reordering in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        if (VB or VBG or VBZ or VBN or VBP) in w_j then
            tmp = w_j
            w_j = w_{j+1}
            w_{j+1} = tmp
            store s_i in S
        end if
    end for
end for
Write S to file
```

B. Reordering rule for question sentence

1. "Questions without question words" reordering rule

Making correctly formed questions without question words in English language is easy, almost all questions use the same structure. All you need to do is put the main verb before the subject. This kind of question is sometimes called Yes/No questions. Yes or no question is a question whose expected answer is either "yes" or "no". On the other hand this is not true for Amharic language because in Amharic the main verbs follow the Object with the normal sentence structure of Amharic language.

Example 4.7

Sentence: You(S) are(V) from Germany.

Question: Are(V) you(S) from Germany?

Example 4.8

Sentence: *Ante(S) ke Germany(O) neh(V).*

Question: አንተ(S) ከጀርመን(O) ነህ(V) ?

As we see in the above Example 4.7 the two sentences have different word order. In the normal sentence structure the subject comes before the main verb, whereas in the question the subject comes after the main verb. But this is not true for Amharic language because the normal sentence and the question have the same word order as we see in Example 4.8, only the intonation will change for command.

Therefore, in order to avoid such word order difference we have reordered the English sentence to the same as the Amharic sentence structure. **Algorithm 6** is used for reordering this type of sentences. After applying the reordering rule the above English question will be as follows:

Question: you(S) Are(V) from Germany?

Algorithm 6 Algorithm for Questions with question words reordering in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        if (VB or VBG or VBZ or VBN or VBP) in w_j and "?" in w(length(s)-1) then
            tmp = w_j
            w_j = w_{j+1}
            w_{j+1} = tmp
            store s_i in S
        end if
    end for
end for
Write S to file
```

2. "Questions with question words" reordering rule

The only difference between question with question words and question without question words is WH question words. Otherwise both question types follow the same sentence structure. The main verb always comes before the subject and the question word always comes before the main verb. However, this is not true for Amharic language. In Amharic the main verb always come after the question word or the object and the subject always come before the question word as we see in Example 4.9.

Example 4.9

Who(QW) is(V) that student(S)? => ያ ተማሪ(S) ማን(QW) ነው(V) ?

As we see in Example 4.9 the two sentences have different word order. So, in order to avoid such word order difference we reordered the English sentence to the same as the Amharic sentence structure. **Algorithm 7** is used for reordering this type of sentences. After applying the reordering rule the above English question will be as follows:

That student(S) who(QW) is(V) ? => ያ ተማሪ(S) ማን(QW) ነው(V) ?

Algorithm 7 Algorithm for Questions without question words reordering in English sentences

Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each $s_i \in \{S\}$ where i =1,2,3...k **do**
    for each $w_j \in \{s_i\}$ where j=0,1,2,3...l **do**
        if (VB or VBG or VBZ or VBN or VBP) in w_j **and WDT or WP** in w_{j+1} **and** "?" in *w(length(s)-1)* **then**
            $tmp = w_j$
            $w_j = w_{j+1}$
            $w_{j+1} = tmp$
            store s_i in S
        end **if**
    end **for**
end **for**
Write S to file

C. Reordering rule for Compound-Complex sentence

A compound complex sentence contains at least two independent clauses and at least one subordinate clause. This clauses are joined by the coordinating conjunction or semicolon. A coordinating conjunction is a word that glues words, phrases and clauses together.

Example 4.10

Tom cried(IC) and I apologized him immediately(IC), because the ball hits him(DC).

In the above Example 4.10, the sentence is compound-complex sentence. The sentence has two independent clauses(IC) and one dependent clause(DC). We can split the sentence in to two simple sentence by using comma. Then the rules which are mentioned for simple sentence, are applied on it. To do this **Algorithm 8** is used.

Algorithm 8 Algorithm for Complex sentence reordering in English sentences

```
Load the sentences from POS Tagged corpus
Store all sentences in S
Store all words in W
for each s_i ∈ {S} where i =1,2,3...k do
    for each w_j ∈ {s_i} where j=0,1,2,3...l do
        if "," in w_j then
            Divide s in to simple sentence as s_1,s_2,...s_n
            Apply reordering rule for simple sentence
            Concatenate s_1,s_2,...s_n
            store s_i in S
        end if
    end for
end for
Write S to file
```

4.2 Decoding

The main part of SMT is *Decoding*. This part has three components **Language model**, **Translation model and the Parallel Corpora**. The language model helps the translation system in selecting words or phrases appropriate for the local context and combining them in a sequence with better word order. The translation model assigns a probability that a given source language sentence generates target language sentence. The parallel corpora is a corpus from both English and Amharic language.

SMT is typically formulated under the framework of noisy channel model. Given a source sentence (in English), E= e1, e2, e3, e4..., we want to find the best Amharic translation A= a1, a2, a3, a4..., among all possible translations as shown in Equation 4.1 (Knight, 1999):

$$A^* = argmax_a P(A|E) \tag{4.1}$$

Where the *argmax* operation denotes the decoder, i.e., the search algorithm used to find among all possible target sentences, the one with the highest probability (Knight, 1999). The above **Noisy Channel Model** is defined as follows. Applying Bayes' decision rule and dropping the constant denominator on Equation 4.2, we have Equation 4.3.

$$P(A|E) = P(E|A) * P(A)/P(E) \tag{4.2}$$

Where E is the source text and A is the target.

$$argmax_a P(A|E) = argmax_a P(E|A) * P(A)/P(E) \tag{4.3}$$

Since E is fixed, we omit it from Equation 4.1 and 4.2, we get the Noisy Channel Model equation, which is:

$$A^* = argmax_a P(E|A)P(A) \tag{4.4}$$

Where $P(A)$ is the language model, and $P(E|A)$ is the translation model, modeling the

transformation probability from E to A.

4.2.1 Language Model

The main aim of language model is to help the translation system with selecting words or phrases appropriate for the local context and combining them in a sequence with better word order. The most common approach used in language modeling is to estimate the probability of a word conditioned on a window of preceding words called the history. These types of language models are called **N-gram language model** (Knight, 1999).
The most widespread statistical language model N-gram model, was proposed by Jelinek and Mercer (Axelrod, 2006) and has proved to be simple and robust. Much like phrase-based statistical machine translation, the N-gram language model has dominated the field since its introduction despite disregarding any inherent linguistic properties of the language being modeled (Axelrod, 2006).
The N-gram model makes the assumption that the n^{th} word **w** depends only on the history **h**, which consists of the $n-1$ proceeding words. By neglecting the leading terms, it models as a Markov chain of $n-1$ order as shown in Equation 3.3 (Axelrod, 2006):

$$Pr(w \mid h) \triangleq Pr(w_i \mid w_1, w_2, ..., w_{i-1}) \approx Pr(w_i \mid w_{i-N+1}, ..., w_{i-1}) \tag{4.5}$$

The value of n trades off the stability of estimate (i.e. its variance) against its appropriateness (i.e. bias). A high n provides a more accurate model, but a low n provides more reliable estimates. Despite simplicity of N-gram language model, obtaining accurate N-gram probabilities can be difficult because of data sparseness. Given infinite amounts of relevant data, the next word following a given history can be reasonably predicted with just the maximum likelihood estimate (MLE) (Axelrod, 2006).

$$P_{MLE}(w \mid h) = \frac{(Count(h, w))}{(Count(h))} \tag{4.6}$$

Where h means a number of words in training corpus.

The count function simply measures the number of times something was observed in the

training corpus. If the value of n increases, direct MLE computation of N-gram probabilities from counts is likely to be highly inaccurate (Axelrod, 2006). Various smoothing techniques have been developed to compensate for the sparseness of N-gram counts, thereby improving the N-gram model's estimates for previously unseen word sequences (Knight, 1999).

Example 4.11, If we take the following Amharic sentences as one simple corpus,

መልካም የገና በዓል ይሁንላችሁ ::

መልካም አዲስ ዓመት ::

መልካም የልደት በዓል ::

መልካም በዓል ይሁንላችሁ ::

መልካም በዓል ለሁላችሁም ::

The **uni-gram** probability can be computed as:

$$P(w_1) = \frac{count(w_1)}{count(h)} \tag{4.7}$$

If we take the word "መልካም" the **uni-gram** probability will be:

$$P(\text{መልካም}) = \frac{5}{16} = 0.3125 \tag{4.8}$$

The **bi-gram** probability can be computed as:

$$P(w_2 \mid w_1) = \frac{count(w_1 w_2)}{count(w_1)} \tag{4.9}$$

If we take the words "መልካም" and "በዓል" the probability will be:

$$P(\text{በዓል} \mid \text{መልካም}) = \frac{2}{5} = 0.4 \tag{4.10}$$

The **trigram** probability can be computed as:

$$P(w_3 \mid w_2 \mid w_1) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)} \tag{4.11}$$

If we take the words "መልካም", "በዓል" and "ይሁንላችሁ" the probability will be:

$$P(\text{ይሁንላችሁ} \mid \text{መልካም} \mid \text{በዓል}) = \frac{1}{2} = 0.5 \tag{4.12}$$

In this research, since the system is unidirectional a language model has been developed for Amharic only by using N-gram language model with the help of IRSTLM tool. IRSTLM is a tool used for developing a language model for machine translation (Tripathi, 2010).

4.2.2 Translation model

The other component in decoding process is translation model assigns a probability that a given source sentence generates target sentence (Osborne, 2011). Translation model is the way sentences in E (English) get converted to sentences in A (Amharic) which is denoted by $P(A|E)$ means probability of A given E. It is calculated as:

$$P(A \mid E) = \frac{(Count(A|E))}{(Count(E))} \tag{4.13}$$

In general, translations require many-to-many alignments between words. This means that group of words in the source sentence should be translated to group of words of target sentence. The model can be *single-word-based translation (SWBT) or phrase-based translation (PBT)*. The basic idea of single-word-based translation (SWBT) is to segment the given source sentence into words, then to translate each words and finally compose the target sentence from these word translations. One major disadvantage of the single-word based (SWBT) approach is that contextual information is not taken into account. which means the lexicon probabilities are based only on single words.

One way to incorporate the context into the translation model is to learn translations for whole phrases instead of single words. Here, a phrase is simply a sequence of words. So

the basic idea of phrase-based translation (PBT) is to segment the given source sentence into phrases, then to translate each phrase and finally compose the target sentence from these phrase translations as shown in Example 4.12. For this study we used phrase based translation model (Gao, 2011).

Example 4.12

If the translation model is **SWBT**

Source sentence: His car is comfy

Target sentence: የሱ መኪና ይመቻል ነው

If the translation model is **PBT**

Source sentence: His car is comfy

Target sentence: የሱ መኪና ይመቻል

The phrase based translation had better translated output compared with the single word based translation output.

In general, the above reordering rules are collected manually from different grammar books of English and Amharic languages and the algorithms used in this section is designed manually.

4.3 Post-processing

Finally Post-processing is used for checking the output of SMT to improve the final output of the system. If the output sentence is grammatically correct the target sentence will be delivered otherwise the grammar corrector corrects the grammar and delivers the corrected sentence. This processing stage has two main components the *Grammar checker and Grammar corrector*.

4.3.1 Grammar Checker

The grammar checker module is used to check whether the translated sentence (the output of SMT system) is grammatically correct or not. The module has five main sub modules **Dictionary, Morphological analyzer, Grammar rule checker, Grammatical Relation Finder, and Language model** (Temesgen, 2013).

Dictionary, the first module in grammar checker. The main purpose of the dictionary is to translate words that are not translated by the SMT decoder. Sometimes the output of the SMT system deliver English words in combination with Amharic words within one sentence as shown in Example 4.13 . This happens because of word limitations in parallel corpora. Making all words in Amharic will help the Grammar checker to take the translated sentence as Amharic Sentence.

Example 4.13

If we take the following sentences

English Sentence: Abebe will come to Addis Ababa for his honeymoon.

The output of the SMT might be :

Amharic Sentence: አበበ ለ honeymoon ወደ አዲስ አበባ ይመጣል

If we take the above Amharic sentence the word *honeymoon* is not translated to its equivalent Amharic word so, we have to use English-Amharic Dictionary for the translation purpose. After getting the equivalent translation from the dictionary the above Amharic sentence will be as follows:

Amharic Sentence: አበበ ለ ጫጉላ ሽርሽር ወደ አዲስ አበባ ይመጣል

Morphological analyzer, module accepts the output sentence (Amharic sentence) from the decoder and assigns linguistic meanings to each word. Each affix (prefix, infix, and suffix) or morpheme in word refers to different linguistic meaning such as number, person, gender, definiteness and the like. Morphological analyzer should be used to know what linguistic meaning that a given affix refers to. In this study we used HornMopho Amharic morphological analyzer which is freely available for Amharic, Tigrigna and Afan Oromo languages (Gasser, 2011).

HornMorpho recognizes two POS tags i.e. noun and verb. If the word is a noun, linguistic meanings such as number, person, gender that noun belongs will be determined using the HornMorpo. This will be done for both subject and object. On the other hand, the verb should be analyzed for both subject and object markers because number, person and gender of the subject and object of the sentence are reflected on the verb as well. Therefore, when the verb analysis is done using HornMorpho the subject and object marker on the verb will be defined in terms of number, person and gender (Gasser, 2011). For example: if we give a word *"ይመጣል"* the morphological analyzer give an output of:

Word: *"ይመጣል"*

POS: verb, root: <mT'>, citation: *መጣ*

Subject: 3, sing, masc

Grammar: imperfective, aux:alle

POS: verb, root: <mTT>, citation: *መጠጠ*

Subject: 3, sing, masc

Grammar: imperfective, aux:alle

POS: verb, root: <mT'>, citation: *ተመጣ*

Subject: 3, sing, masc

Grammar: imperfective, aux:alle, passive

Grammatical relation finder

Grammatical relation finder module assigns grammatical relation between words in a sentence such as the subject and verb, the object and verb and so on. The grammatical relation is constructed based on the analysis result of the Morphological analyzer. Grammatical relation finder assigns subject-verb and object-verb relation in terms of number, person and gender. The relation between the words is represented in the form of word pattern as shown in Example 4.16 (Temesgen, 2013).

Example 4.14

If we take this sentence

Amharic Sentence: **እሱ ተማሪ ነው**

The sentence could be put in three different word patterns for **Number**, **Person** and **Gender**. The following three lines show how Grammar Relation Finder generates sentence pattern for the given sentence. In terms of number, person and gender respectively.

('SStart NSng NSng VSng SEnd', 'SStart NP3 NP3 VP3 SEnd', 'SStart NM NN VM SEnd')

Where:

- **SStart** shows the beginning of the sentence.
- **NP1/NP2/NP3** shows the word is first person , second person or third person.
- **NM/NF/VM/VF** shows the word(Noun, Verb) is Masculine or Feminine.
- **NN** shows if the word can not be classified by gender.
- Finally **SEnd** shows the end of the sentence.

No	Common grammar errors
1	Subject - verb disagreement
2	Object - verb disagreement
3	Incorrect Word order
4	Adjective - noun agreement
5	Adverb - verb agreement

Table 4.4: Common grammar errors of Amharic language

```
<rules>
<rulegroup id="Number">

<rule id="_NpVSbS" name="Subject/Verb agreement in number(subject plural and verb singular)">
<pattern >SStart NPlr (NPlr|NSng) VSbSng (VObPlr|VObSng|VObNN) SEnd</pattern>
<message>ለብዙ ቁጥር ድርጊት ፈፃሚ ብዙ ቁጥር ድርጊት ፈፃሚ የሚያሳይ ግስ ተጠቀም</message>
<example type="correct">በሬዎቹ ልጅቷን ወጓት</example>
<example type="incorrect">በሬዎቹ ልጅቷን ወጋት</example>
</rule>

<rule id="_NsVSbp" name="Subject/Verb agreement in number(subject singular and verb plural) ">
<pattern >SStart NSng (NPlr|NSng) VSbPlr (VObPlr|VObSng|VObNN) SEnd</pattern>
<message>ለነጠላ ቁጥር ድርጊት ፈፃሚ ነጠላ ቁጥር ድርጊት ፈፃሚ የሚያሳይ ግስ ተጠቀም</message>
<example type="correct">በሬው ልጅቷን ወጋት</example>
<example type="incorrect">በሬው ልጅቷን ወጓት</example>
</rule>

</rulegroup>
```

Figure 4.2: XML file format for Language model

Language model is collection of Grammar rules for Amharic based on Gender, Number and Person. Grammar rules in the model are manually constructed rules. Those rules describes what must not occur in Amharic sentence or incorrect grammatical sentence. If a sentence matches with any of the rules then it will be considered as a grammatically incorrect sentence (Temesgen, 2013). The rules are written in regular XML format. In this study we used common grammar errors of Amharic language which are listed in Table 4.4 to generate the language model in XML file.

In the XML file (Language model) the root element for the rule set is **"rules"**. The root element **"rules"** has two children named **"rule"** and **"rulegroup"**. The rule has three sub elements: **"pattern"**, **"message"** and **"example"**. There is also **"rulegroup"** set of rules with rulegroup id for each Grammar error type(Gender, Number and Person errors). We can see how those rules are modeled in Figure 4.2.

Grammar rule checker, module matches the patterns extracted in the grammatical relation finder module against the rules in the language model. If the extracted pattern from the grammar relation finder of the sentence matches against the pattern in the language model, the sentence will be considered as grammatically incorrect. Otherwise, the sentence would be considered as correct.

Example 4.15

If we take the this sentence

Amharic Sentence: **እሷ ደሃ ነው**

The three sentence pattern generated by Grammar relation finder for the above Amharic sentence are:

('SStart NSng NSng VSng SEnd', 'SStart NP3 NP3 VP3 SEnd', 'SStart NF NN VM SEnd')

This sentence has a problem on gender, because for feminine subject it uses masculine verb as we see in the third pattern. The grammar checker identified this kind of problem and send the error for grammar corrector to correct the sentence.

4.3.2 Grammar Corrector

The last component in Post-processing is grammar corrector. The corrector is used for correcting grammatically incorrect sentence which is identified by the Grammar checker. The grammar corrector component has grammatically correct rule patters like language model in grammar checker. The module accepts the output sentence of the grammar checker with its error type. Then put the sentence with the correct pattern and provide the corrected sentence. This component has three special types of Amharic-Amharic dictionaries *Gender based Dictionary*, *Person based Dictionary* and *Number base Dictionary*. The dictionaries are used for correcting the grammatically incorrect sentence in

each category. If we take sentence in Example 4.15, the corrected sentence is:

Amharic Sentence: አሷ ደሃ ነች

The word "ነች" is extracted from gender dictionary. The above three dictionaries are manually constructed dictionaries. The dictionaries used in grammar corrector component are limited in size. This might have an impact on the final result. A more comprehensive dictionary could give a better result.

You can find all python and XML codes used for reordering, grammar checker and corrector in the this url: `https://goo.gl/ZjhvGk`

Chapter 5

Evaluation

In this Chapter, we investigate if the proposed hybrid approach improves the performance of existing state of the art SMT based Amharic-English translator. In particular, we try to answer the following research questions (RQ)s.

RQ1: "Is hybrid MT approach better than SMT approach for English−Amharic MT system?"
RQ2: "Which RBMT component has better contribution to the hybrid system, post-processing or pre-processing?"

RQ1 is raised to investigate if hybrid based English-Amharic MT system can improve the translation accuracy of SMT based English-Amharic MT system. This investigation is needed to see if the added components in the proposed hybrid approach helps to improve the performance of the English-Amharic translation system. The second research question RQ2 is raised to identify which rule based component has more impact on performance improvement of English-Amharic translation system.

5.1 Procedure

To answer the first research question we are going to use the following process:

1. Reproduce SMT system for English-Amharic.
2. Develop hybrid system for English-Amharic.
3. Implement and test both systems independently.
4. Compare the result of both systems and finally conclude.

To answer the second research question we are going to follow the following process:

1. Develop SMT system for English-Amharic.
2. Apply and test hybrid system without post-processing.
3. Apply and test hybrid system without pre-processing.
4. Compare the result of the above two hybrid approaches with SMT approach one by one.

5.2 Software tools

We used several machine translation tools for our Hybrid approach MT system, Moses, GIZA++, IRSTLM, Python, Horn morpho and Stanford POS tagger.

MOSES : Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair. All you need is a collection of translated texts (parallel corpus). Once you have a trained model, an efficient search algorithm, it quickly finds the highest probability translation among the exponential number of choices (Koehn, 2017).

GIZA++ : The translation table was produced by GIZA++, which trains phrase-based translation models from aligned parallel corpora (English-Amharic). GIZA++ induces phrase-level alignments between sentences (Koehn, 2017).

IRSTLM : Language Modeling Toolkit, features algorithms and data structures suitable to estimate, store, and access very large N-gram language models (Koehn, 2017).

PYTHON : is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together.

HORN MORPHO : HornMorpho is a set of Python programs for analyzing and generating words in Amharic, Tigrinya, and Afan Oromo. A user interacts with the programs through the Python interpreter (Gasser, 2011).

STANFORD POS TAGGER : software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc (Jurafsky and Martin, 2014).

5.3 Parallel Corpora

A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages than the original. We collected English-Amharic text Parallel corpus for the MT system. There are two type of text corpus: Simple sentence corpus and Complex sentence corpus. The corpuses are collected for both Amharic and English languages. The size of the corpuses are shown in Table 5.1. The Complex sentence corpus is collected from *King James Version(KJV)* of English-Amharic Holy bible and the simple sentence is collected from previous researchers. We have to consider that the training corpus of simple sentence and complex sentence are different. That means the training and testing is done independently.

Parallel Corpus		
Sentence type	**Amharic**	**English**
Simple sentence	3915	4960
Complex sentence	25218	46933

Table 5.1: Parallel corpus size in words

5.4 Evaluation method

Evaluation of machine translation systems can be done either by *Manual Evaluation method (Human evaluation) or Automatic Evaluation method (Software based).* For this study we used software called BLEU score which is based on automatic metrics.

Manual evaluation

Manual evaluation is done based on their fluency as well as the adequacy of their content (Axelrod, 2006). Fluency, answers the question "is the output good fluent or not?" and adequacy answers the question "Does the output convey the same meaning as the input sentence?". Manual translation is time-consuming and expensive to compare different versions of a system during development and testing.

Automatic evaluation

Automatic evaluation is one of the most crucial issues in the development stage of a MT system, given that manual evaluation are usually expensive. Error rate is typically measured by comparing the system output against a set of human references, according to an evaluation metric of choice. By far, the most widely used metric in recent literature is BLEU(Bilingual Evaluation Understudy) (Papineni et al., 2002).

BLEU score is defined in the range between 0 and 1 (or in percentage between 0 and 100), 0 meaning the worst translation (where the translation does not match the reference in any word), and 1 the perfect translation. BLEU score computes lexical matching accumulated precision for N-grams up to length four (Papineni et al., 2002).The central idea behind the metric is that "the closer a machine translation is to a professional human translation, the better it is".

BLEU score uses a modified form of *precision* to compare a candidate translation against multiple reference translations. The metric modifies simple precision since machine translation systems have been known to generate more words than appear in a reference text. No other machine translation metric is yet to significantly outperform BLEU with respect to correlation with human judgment across language pairs. We can see how BLEU score

is defined in Equation 5.1.

$$BLEU score = min(1, \frac{outputlength}{referencelength})(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} \quad (5.1)$$

Where *output length* is a length of machine translation output *reference length* is a length of human reference translation. The precision is computed over the entire corpus, not single sentences for each N-gram models. Example 4.13 shows how each how each n-gram precision is calculated.

Other evaluation metrics

Apart from these, several other automatic evaluation measures comparing hypothesis translations against supplied references have been introduced, claiming good correlation with human intuition. Although not used in this study, here we refer to some of them (Clemente, 2008).

- Geometric Translation Mean, or **GTM**, measures the similarity between texts by using a unigram-based F-measure.
- Weighted N-gram Model, or **WNM**, is a variation of **BLEU** which assigns different value for different N-gram matches.
- **ORANGE** uses unigram co-occurrences and adapts techniques from automatic evaluation of text summarization, as presented in the **ROUGE** score.
- As a result from a 2003 John Hopkins University workshop on Confidence Estimation for Statistical MT, introduce evaluation metrics such as Classification Error Rate **(CER)** or the Receiving Operating Characteristic **(ROC)**

The main goal *Automatic Evaluation* is that it computes the quality(accuracy) of translations and compute similarity between the output(Machine translation output) and its reference(original document). It is consistent, tunable and cheap when we compare with Manual Evaluation method (Human evaluation) (Axelrod, 2006).

Chapter 6

Result and discussion

This chapter gives a detailed explanation about experimental result and discussion of English-Amharic machine translation system using Hybrid approach and using SMT approach.

6.1 Result

Eight experiments are conducted to answer RQ1 and RQ2 (see Tables 6.1 and 6.2). The results are based on BLEU score evaluation on moses decoder (Koehn, 2017).

Table 6.1 shows the experimental result of an English-Amharic MT system using SMT approach and Hybrid approach on simple and complex sentences. These experiments are prepared to compare the two approaches (hybrid and statistical English-Amharic MT system) and to answer RQ1. Both approaches are tested on the same parallel corpus (see, Table 5.1). The first row in Table 6.1 shows accuracy result obtained from English-Amharic MT system using SMT approach only for simple and complex sentences. These experiments are used as a baseline to compare with the result of the hybrid approach. The second row in Table 6.1 shows accuracy result obtained from the hybrid system, by applying both pre and post processing stages. From the evaluation result, the proposed hybrid based English-Amharic MT system achieved an improvement of 14.45% for simple sentence and 20.31% for complex sentence over SMT. ***This is because the effects of pre and post processing (reordering the structure of English sentence structure to Amharic sentence structure and grammar corrector) on inputs and outputs of SMT system.***. The hybrid approach gives more accurate English to Amharic Translation than SMT.

BLEU Score result		
Approach	**Simple sentence**	**Complex sentence**
SMT	30.67%	26.61%
HMT	**45.12%**	**46.92%**

Table 6.1: Result of English-Amharic MT system using SMT and HMT

BLEU Score result		
System	**Simple sentence**	**Complex sentence**
SMT	30.67%	26.61%
HMT using only pre-processing	33.79%	**46.61%**
HMT using only post-processing	**39.87%**	29.65%

Table 6.2: Result of English-Amharic MT system using SMT, HMT using only pre-processing and HMT using only post-processing

Table 6.2 shows the experimental result of an English-Amharic MT system using SMT approach, hybrid approach based on only pre-processing and hybrid approach based only post processing on simple and complex sentences. These experiments are prepared to investigate which rule based component has more contribution to the hybrid system, this answers the research question raised on RQ2. The approaches are tested on the same parallel corpus (see, Table 5.1). The first row in Table 6.2 shows accuracy result obtained from English-Amharic MT system using SMT approach only for simple and complex sentences.

From the evaluation result, the proposed hybrid English-Amharic system based on only pre-processing stage achieved an improvement of 3.12% and 20% for simple sentence and complex sentences respectively over SMT based system. The proposed hybrid English-Amharic system based on only post-processing stage achieved an improvement of 9.2% and 3.04% for simple sentence and complex sentences respectively over SMT based system. ***This shows the effects of pre-processing (reordering English sentence structure to Amharic sentence structure) on inputs of SMT results in higher improvement for complex sentences and the effects of post processing (Grammar checker and corrector) on outputs of SMT results in higher improvement for simple sentences.***

6.2 Discussion

Table 6.1 shows the experimental result of English Amharic MT system using *SMT approach* and *hybrid approach* on simple and complex sentences. This experiments are prepared to answer RQ1 "Is hybrid MT approach better than SMT approach for English-Amharic MT system ? ". To answer the question, we compared the result of SMT and HMT for both simple and complex sentence. The result shows HMT using pre and post processing achieved almost 20% and 15% accuracy over the SMT system for complex and simple sentences respectively. The improvement is achieved due to the effect of adding linguistic knowledge of both Amharic and English languages on the SMT system. This linguistic knowledge is used to reorder syntactical structure of English sentence to Amharic sentence before forwarding the input sentence to SMT decoder as well as to develop rule based Amharic grammar checker to post process the outputs of SMT decoder. This helps the translation system to generates more perfect translated target sentence. This means the translated sentence is close to human translation.

Table 6.2 shows the experimental result of an English, Amharic MT system using *SMT approach*, *hybrid approach based on only pre-processing* and *hybrid approach based only post processing* on simple and complex sentences. The experiments are conducted to answer RQ2 "Which RBMT component has better contribution to the hybrid system ?", we performed two comparison the first one is , the experimental result of SMT with HMT using only pre processing and the second one is SMT with HMT system using only post processing for both simple and complex sentence.

On the first comparison *hybrid approach based on only pre-processing* achieved higher accuracy than *SMT*. But as we see from the result for simple sentence the improvement is not much as compared to complex sentence. This is because simple sentences are short in length(contains 3-5 words in a sentence) this decreases difficulties of reordering for SMT system. But this is not true for complex sentence because complex sentence reordering needs more linguistics knowledge on both source and target languages.

On the second comparison *hybrid approach based only post processing* achieved higher accuracy than *SMT*. But as we see from the result of BLEU score, simple sentence improvement is better than complex sentence. This shows post processing has more contribution on simple sentences. This is because the main aim of post processing is to check and correct grammatical errors based on gender, person and number. In this research we used Holy Bible document for complex sentence. Such those grammatical errors are rare in Biblical document. This is the reason why complex sentence improvement is less than simple sentence. In general, research question two investigates the effects of *pre/post processing* on English-Amharic MT system.

6.3 Threat to validity

Internal threats to validity

This kind of threat is caused by variation in instrumentation, and effect due to uncontrolled variables. The instrumentation used to measure variables is computer. We run all experiments on the same computer to avoid instrumental variation. The effect of N-gram model generates zero probability, which happen when we measure the accuracy on BLEU score measurements. N-gram models can assign non-zero probabilities to sentences they have never seen before. The only way you'll get a zero probability is if the sentence contains a previously unseen bigram or trigram. In that case, we can do smoothing technique, i.e adding smoothing coefficient to N-gram precision that have zero probability.

External threats to validity

The proposed hybrid approach is evaluated based on some specific domain dataset and the evaluation is conducted using only automatic MT evaluation method. Our proposed approach is independent of the data and evaluation metrics. The approach could be used and evaluated on any other English-Amharic dataset, which could have large amount of parallel corpora for English-Amharic languages. The proposed MT system could also be evaluated using a manual evaluation method.

Conclusion threat to validity

The main threats to conclusion validity is selecting test data. We used split sampling and we didn't use cross validation technique for testing the performance of MT system. Cross validation is not used because we prepared four kinds of dictionary manually for grammar checker discussed in section 4.2.3. Those dictionaries are not available. In in order to solve this problem we prepared the dictionaries for words in test data only. Our proposed approach could be tested if there is any available large size dictionaries.

Chapter 7

Conclusion and Future work

7.1 Conclusion

This study tried to address the problem of English-Amharic Machine translation system. We proposed English-Amharic MT system using a hybrid approach. We compare our proposed approach with the existing English-Amharic MT system that uses SMT approach. In particular, we investigated, if the proposed approach has better accuracy than SMT approach. We also further analyzed the hybrid system to identify which hybrid component contributes to the improved performance observed. We evaluate the proposed approach using BLEU score evaluation method.

The evaluation shows that the proposed hybrid approach is better than the existing SMT approach for English-Amharic MT system. We got 20% and 15% accuracy improvement for complex and simple sentences respectively. We observed that integrating pre-processing on SMT system has effect on the efficiency of the translated output of SMT for complex sentences. Integrating post processing on outputs of SMT will have more effect on accuracy of the simple sentence translation.

7.2 Future work

The main limitations of this research are lack of documented parallel corpus for English-Amharic languages, non-availability of *Person, Number* and *Gender* based dictionaries for Amharic. We were forced to use bible corpus for complex sentences and some collected simple sentence corpus from previous researchers. We developed our own simple dictionaries for the grammar checker based on *Person, gender* and *Number*. Based on the aforementioned limitations of this research, we recommend the following points to be addressed as a future work.

- The proposed approach should be tested on large data sets.
- Prepare domain based large size dictionaries in Amharic language and apply on the proposed approach for better improvement.

Bibliography

Axelrod, A. (2006). Factored language models for statistical machine translation.

Baye, Y. (2000). *Amharic Grammar*. Berahana selam.

Clemente, J. M. C. (2008). Architecture and modeling for n-gram-based statistical machine translation. *Universitat Politecnica de Catalunya, Spain*.

Costa-Jussa, M. R., Farrús, M., Marino, J. B., and Fonollosa, J. A. (2012). Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and informatics*, 31(2):245–270.

Costa-Jussa, M. R. and Fonollosa, J. A. (2015). Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1):3–10.

Daba, J. (2013). *Bidirectional English–Afaan Oromo Machine Translation Using Hybrid Approach*. PhD thesis, Addis Ababa University.

España Bonet, C., Màrquez Villodre, L., Labaka, G., Díaz de Ilarraza Sánchez, A., and Sarasola Gabiola, K. (2011). Hybrid machine translation guided by a rule-based system. In *Machine translation summit XIII: proceedings of the 13th machine translation summit, September 19-23, 2011, Xiamen, China*, pages 554–561.

Gao, J. (2011). A quirk review of translation models.

Gasser, M. (2011). Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*.

Gasser, M. (2012). Toward a rule-based system for english-amharic translation. *Language Technology for Normalisation of Less-Resourced Languages*, page 41.

Getahun, A. (2010). *Modern Amharic Grammar in a simple approach*. Berahana selam.

Hutchins, W. J. (1995). Machine translation: A brief history. *Concise history of the language sciences: from the Sumerians to the cognitivists*, pages 431–445.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*. Pearson.

Kim, J.-B. and Sells, P. (2008). *English syntax: An introduction*. CSLI publications.

Knight, K. (1999). A statistical mt tutorial workbook.

Koehn, P. (2017). Machine translation system user manual and code guide.

Okpor, M. (2014). Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159.

Osborne, M. (2011). Statistical machine translation. In *Encyclopedia of Machine Learning*, pages 912–915. Springer.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Piscioneri, M. et al. (2011). English grammar today, reference book and workbook [book review]. *English Australia Journal*, 27(1):87.

Slocum, J. (1984). Machine translation: its history, current status, and future prospects. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pages 546–561. Association for Computational Linguistics.

Tadesse, A. (2012). English to amharic machine translation using smt.

Temesgen, A. (2013). *Design and Development of Amharic Grammar Checker*. PhD thesis, Addis Ababa University.

Teshome, E. (2013). *Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus*. PhD thesis, Addis Ababa University.

Teshome, M. G. and Besacier, L. (2012). Preliminary experiments on english-amharic statistical machine translation. In *SLTU*, pages 36–41.

Thurmair, G. (2004). Comparing rule-based and statistical mt output. In *The Workshop Programme*, page 5.

Tripathi, e. (2010). Approaches to machine translation.