International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-15 Issue-2, May 2025

Bidirectional English to Wolaytta Machine Translation Using Hybrid Approach

Elisaye Bekele Milke, Tibebe Beshah Tesema, Mesfin Leranso Betalo

***Abstract:** As a part of natural language processing (NLP), machine translation focuses on automated techniques to produce target language text from the source language text. In this study, we combined two approaches: the rule-based MT approach and the statistical MT approach. Sentence reordering, Language model, Translation models, and decoding comprise the system. POS tagging was used to reorder the sentence more comparably, the IRSTLM tool was used to create language models for English, and the Wolaytta, Giza++ tool was used for translation. To ensure mutual translation, two language models have been developed. Four phases of experiments are carried out on the collected data set. Phases of experimentation include preprocessing on the parallel corpus, language modeling, training the translation model, and tune-up the translation system. For both side translations, the BLEU score assessed the accuracy of the translation from Wolaytta to English was 46.31 % and from English to Wolaytta was 56.56%.*

***Keywords:** Natural Language Processing, English-Wolaytta Machine Translation, Machine Translation, bidirectional Machine Translation, Hybrid Approach, Statistical Approach, Rule-Based Approach, Parallel Corpus.*

I. INTRODUCTION

A language is a structured communication system used by humans that consists of sounds (spoken languages) or gestures (sign languages) [1]. A natural language is any language developed naturally in humans through use and repetition without consciousness and deliberate planning. Natural language can take various forms, such as speech or signs. English is the language of the United States, the United Kingdom, Canada, Australia, Ireland, New Zealand, and numerous island states in the Caribbean and the Pacific Oceans, and is native to England. Most of the material, software, and other applicable literature are in English only [2]. The Omotic languages are a group of about 30 languages spoken in southwest Ethiopia around the Omo River, including; the 28 Omotic languages are divided into northern and southern subfamilies [3].

Manuscript received on 12 December 2024 | First Revised Manuscript received on 21 December 2024 | Second Revised Manuscript received on 16 March 2025 | Manuscript Accepted on 15 May 2025 | Manuscript published on 30 May 2025.
*Correspondence Author(s)

Elisaye Bekele Milke*, Department of Information Technology, Wolaita Sodo University, Ethiopia. Email ID: elisaye.bekele@wsu.edu.et, ORCID ID: 0000-0003-1553-338X

Tibebe Beshah Tesema, Department of Information System, Addis Ababa University, Ethiopia. Email ID: tibebe.beshah@aau.edu.et, ORCID ID: 0000-0001-6418-0707

Mesfin Leranso Betalo, Department of Information Technology, Shenzen University, China. Email ID: mesfinleranso@szu.edu.cn, ORCID ID: 0000-0001-7529-0696

The Wolaytta language is one of the northern Omotic languages spoken in the Wolaytta area and some other parts of the nations, nationalities, and peoples region of southern Ethiopia [4].

The Latin script has been used since 1993 to write Wolaytta's texts. The Wolaytta people are one of Ethiopia's indigenous peoples, whose cultures, traditions, political legacy, and kingdom belong to them. Wolaytta belongs to the Omotic-speaking people and their language is called Wolaytta (Wolayttato Doona) after the city's name [5].

The translation of a natural language (source language (SL) to another language (target language (TL) using computer systems with or without human interaction" is defined as machine translation [6]. One of the greatest advantages of machine translation is that it translates large lengths of text in a very short time, enables a quick and comprehensive understanding of the document, and is inexpensive. As a general term, MT is categorized into four main parts, "rule-based" refers to machine translation systems that are built based on both linguistic information about the source and target languages, primarily obtained from dictionaries (bilingual), and "example-based" machine translation with parallel texts as the main knowledge, where the technology is the main idea. As a result of the analysis of bilingual texts, "SMT" is generated by statistical models. Using rule-based, example-based, and statistics-based translation methods, a new approach, called the "hybrid approach", has been developed which has proven to be more efficient in the field of machine translation systems [7].

A. Problem Statement

Artificial intelligence is used in machine translation to translate text from one language to another without requiring a human translator. Bilateral machine translation systems between English and Wolaytta languages have not been created or investigated. Language speakers require human translators to translate various articles, books, and other written materials from English to the Wolaytta language and vice versa. For example, both English and Wolaytta language speakers encounter unique obstacles. However, the drawback of human translators is that they cannot speed up turnaround times and are more expensive than machine translation, which is why machine translation is 100 times more affordable than human translation. A human translator cannot handle a vast volume of swiftly translated text. Human translators cannot alter the cost, quality, and time in a way that was never feasible with linguistic solutions in the past.

However, in conducting this work many challenges affect the bidirectional English-Wolaytta translation system. These challenges are as described below:

1. Finding parallel data sets for low-density language

pairs, in the case of our study, the Wolaytta language has no electronic form of a dataset.

2. Machine translation between language pairs belongs to distinct families or language pairs having different word order (e.g. between SVO and SOV word order languages.)
3. Not all the words in the English language have equivalent words in the Wolaytta language and vice versa. In some cases, a word in the English language is to be expressed by a group of words in Wolaytta and vice versa.
4. The challenge with machine translation is that it cannot account for the subtleties and nuance of language.

II. LITERATURE REVIEW

According to investigated Bidirectional Tigrigna-English Statistical Machine Translation. The investigator used three sets of experiments: baseline (phrase-based machine translation system), morph-based (based on morphemes obtained using an unsupervised method), and post-processed segmented systems (based on morphemes obtained by post-processing the output of the unsupervised segmented). However, the investigator used only a statistical machine translation approach and this approach focuses only on the probability of the word occurring in the corpus. If the collected corpus is little, the accuracy of the translation is also too low [7].

According to investigated Bidirectional, English-Amharic machine translation using constraint corpus. To achieve bi-directional translation, the investigator used two language models were developed, one in Amharic and the other in English. In this investigation Translation models have been developed to assign a chance for a given source language text to produce a target language text. Data collection-based experiments were conducted and results were recorded. The trials were conducted separately, in single sentences and another in complex sentencing. However, the limitation of this study was that the author not considers the Part of speech tag (POST) to get better translation accuracy [8].

According to [9], investigated on English to Wolaytta machine translation using a statistical approach. The investigator used MGIZA++ to align the corpus to the word level by using IBM models and SRLM for the language model. However, the limitation of this study is that the investigator did not consider the accuracy of the translation system used only a statistical approach, and did not identify the structure of both language sentences. Also, the investigated translation system is only one direction.

According to bidirectional Amharic-Afaan Oromo machine translation using a hybrid approach. The developed system includes four components: sentence rearrangement, model language, decoding, and model translation. The investigator used the IRSTLM and GIZA++ toolkit for the language model and translation model. However, the investigator does not consider the rule parts of both languages and the translation accuracy is low [10].

The main research contribution is summarized as follows:

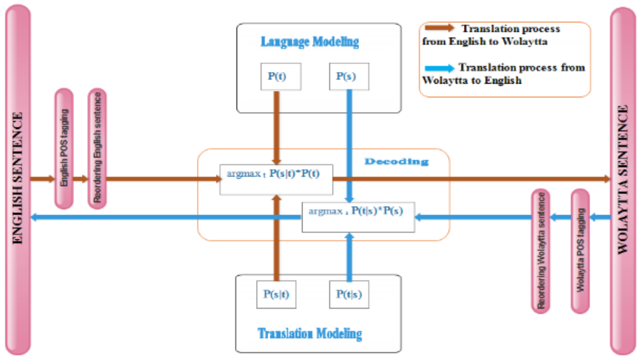
- We enable effective communication between Wolaytta language speakers and English language speakers through a bidirectional machine translation system.
- We Examine and identify key features for bidirectional English-Wolaytta machine translation.
- The developed machine translation system is 100 times less expensive than human translation.
- This work partially fulfills the Information gap of Wolaytta language speakers because the web resources like books, articles, news, journals, and other materials are written in English and other European languages.

III. DESIGN AND DEVELOPMENT OF THE SYSTEM

The study plans to develop a two-way machine translation system in English and Wolaytta languages with a hybrid approach, for the good fulfillment of this study, some measures have been taken, A corpus has been developed based on the approach followed and different tools have been used to develop a working system.

A. Architecture of the System

Figure 1 shows the architecture of bidirectional English-Wolaytta machine translation using a hybrid approach. The architecture is made up of five major components, namely; POS Tagging, Sentence Reordering, Translation model, Language model, and decoder. The components are described as follows:



[Fig.1: Architecture of the System]

B. POS Tagging

This component is the first step in the part of the rule, which assigns speech parts for the English and Wolaytta phrases to each word. A POS tag (or Part of speech tag) is a special level that is assigned to each token (word) in a text corpus to classify the part of speech and often other grammatical categories such as time, number (plural/singular), uppercase / Lower case, etc. POS tags are used for corpus search and text analysis tools and algorithms. In this study, for the English corpus Penn Tree bank Speech Tag Set is annotated, which was developed by Helmut Schmidt in the TZ project at the Institute for Computational Linguistics at the University of Stuttgart and contains modifications developed by Sketch Engine and for Wolaytta corpus there is no pre-made tag set in the Wolaytta language [11]. In this section,
label sets for the study are identified and developed. Identifying and developing the details of the tag set requires human expertise and is time-consuming. Therefore, different categories of tag sets were identified for this study

Retrieval Number: 100.1/ijsce.B102814020125
DOI: 10.35940/ijsce.B1028.15020525
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

based on the basic Wolaytta language class of speech. For example, if the input phrase is “she is my sister,” the POS tagger will assign parts of speech to each word in the sentence and display it as “she_PRP is_VBZ my_PRP$ sister_NN,” or if the input sentence in Wolaytta as “A ta michchiyo,” the POS tagger will assign A_PP ta_PP$ michchiyo_NN.

C. Sentence Reordering

Reordering is a preprocessing stage for the statistical machine translation system (SMT) in which the words in the source sentence are reordered according to the syntax of the target language. English has an SVO sentence structure while Wolaytta has an SOV sentence structure. The reorganization also supports the decoding process and thus improves the quality of the machine translation. Alignment is the predominant approach to decoding and the accuracy of the translated sentence depends on the alignment of the word. Rearrange to produce fluid and equivalent output in the target language that preserves the meaning of the source text. This rearrangement technique is advantageous in minimizing the syntactic rearrangement problem with SMTs. We classify the rules for rearrangements into three main categories: the rules for simple phrases, questions, and complex phrases. There are some expressions and sentences in Wolaytta and English that do not require reordering the rules for translation.

Example 1: A sentence containing the possessive pronoun which is underlined.

- His cows are grazing grass.
- A miizzati maatta miiddidosana

Example 2: A Sentence containing the interrogative word which is underlined.

- What is your name?
- Ne Sunday oone?

Example 3: A Sentence containing prepositions with cardinal numbers.

- With two people
- Na77u asura issue

D. Translation Model

The translation model gives a chance for a particular sentence to be translated into a sentence in the target language in the source language. When the sentence is transferred to the target language, the translation model finds the highest probabilities for sentences in the source language (t). For the translation of this system, two translation models have been developed the System is bidirectional, if the translation goes from the English sentence with the Wolaytta phrase; it is used to determine the translation quality in English of the source phrase to that Wolaytta Target phrase.

The likelihood of a source sentence being translated into a target sentence is determined by a translation model. Modern translation models can be divided into four categories [12].

i. Word-based Models

The word alignment is called when words in the source phrase are mapped to terms in the objective phrase.

S1...sl and T1...tm are aligned by A=a1, a2,.am, where aj=0,1,...l. *l*}

ii. Phrase-based Models

They were revolutionary, but they didn't take into account case, gender, or homonymy. Every word was translated in a single-true way by the machine, according to it. No restrictions apply to word-by-word translations in phrase-based translation systems. An important step was from traditional "word-based" (IBM-style) models. Using phrase-based translation models, researchers have been able to improve translation quality over IBM models [12].

iii. Syntax-based Model

Instead of words or strings, syntax-based translation is utilized when translating syntactic units. As the syntactic structure of one language is distinct from another, it is difficult to translate machines. If you're translating from English to Wolaytta, you'll need to reorder the words.

iv. Factored based Model

Using factored translation models, phrases are represented as a series of inflected words, rather than as a series of inflected words with multiple levels of information. When using a factored model, a single word is a set of factors. These tags can be used to easily integrate morphological information as well as shallow syntax. Pre-processing and post-processing steps can benefit from such information. To improve statistical machine translation, factored models add different information. Both training data and models incorporate this information.

E. Language Model

The language model has the potential for text strings that can be defined as p(s) (for the source phrase S) and p(t) (for each target sentence T). It helps to select and combine the translation system with appropriate words or phrases in the local context. The objective of the language model is to create a statistical language model that can accurately estimate the propagation of natural language [13]. In the language, in question, you agree to rearrange words. The language used is how fluent one is in the target language of the translated text, i.e. assuming the word order in the target language which is probably more fluent. There are several open-source language modeling toolkits: for integration into the Moses SMT system, RILM, IRSTLM, CMU SLM, and KenLM have been developed [14].

A set of tools for creating and using statistical language models is provided in the SRILM toolkit. Unigrams, bigrams, trigrams, or higher n-grams
can be obtained in the model N-Grams. For instance, if we have the Wolaytta phrases,

- Maji boorra shammiis
- Maji boorra bayziis
- Maji ashuwa miis

- Maji laaxa miis
- Ufaysoy laaxa mibena
- Ufaysoy maja isha
- Ufaysoy ashuwa dosenna

❖ The **unigram** probability can be computed as:

$$P(w1) = \frac{count(w1)}{total\ words\ observed} \Rightarrow P(Maji) = \frac{4}{21} = 0.1904$$

❖ The **bigram** probability may be calculated using the following formula:

$$P(w2|w1) = \frac{count(w1w2)}{count(w1)} \Rightarrow P(Maji\,|\,boorra) = \frac{Count(Maji\ boorra)}{Count(Maji)} = \frac{2}{3} = 0.667$$

Where the words "Maji" and "boorra" have been found in the corpus two times and there are the words "Maji" in the corpus three times, the trigram probability is that:

❖ The **trigram** probability may be calculated using the following formula:

$$P(w3|w1w2) = \frac{count(w1w2w3)}{count(w1w2)} \Rightarrow P(shamiis\,|Maji\ boorra) = \frac{Count(Maji\ boorra\ shamiis)}{Count(Maji\ boorra)} = \frac{1}{2} = 0.5$$

Where 1 is the number of occurrences of Maji, Boorra, and Shammiis, and 2 is the number of occurrences of Maji and Boorra combined.

As the system is bidirectional, a language model was developed in the IRSTLM tool and the N-gram language model, both for English and for Wolaytta.

F. Decoding

After building the language and translation model, the decoding seeks the best possible probability among the exponential of choice for the new entry for the translation of a sentence or phrase, the product of the probability of the English Wolaytta translation model P (s|t) and Wolaytta-English language model p (t), i.e.

argmax$_t$ P(s|t)*P(t)………………………………(1)

If the best translation from Wolaytta to English maximizes the product of the Wolaytta probabilities, English translation **model** P (t|s) and Wolaytta language model English should be used P (s), i.e.

argmax$_s$ P(t|s)*P(s) …………………………..…. (2)

IV. EXPERIMENTATION AND ANALYSIS

A. Corpus Preparation

The hybrid approach requires a parallel bilingual corpus. This study requires parallel documents in English and Wolaytta and comes from the Holy Bible, Wolaytta Zone Education Bureau, and Wolaytta Sodo University Department of Language and Literature. The entire corpus for experimenting with the system contains only 3,492 sentences because of the scarcity of electronic documents within the Wolaytta language, which are very small sentences that compare to most of the available document languages.

B. The Parallel Corpus Preprocessing

After the parallel corpus has been created, there are important preprocessing aspects that must be addressed with regard to the properties of the corpus. The preprocessing of the parallel corpus consisted of four steps, including POS Tagging, text tokenization, true casing, and cleaning. All of these preprocessing steps are carried out with the Perl program in Moses scripts, except for POS tagging.

i. Tokenization

Tokenization is the process of breaking up the flow of text into words, phrases, symbols, punctuation marks, and other elements [15]. We tokenize the English-Wolaytta parallel corpus with Moses Tokenizer.perl, which is available in the Moses software package [16].

The output of the program was text.tok.en and text.tok.wl [17].

ii. True Casing and Lower Casing

True case letters are the problem of words that appear in the text in upper and lower case [18]. The word DOG, Dog, and dog can appear in a large collection of text [19]. Word in the sentences and also does the word count, every word in the corpus has been counted and returns the occurrence number. We performed true casing on English-Wolaytta parallel corpus using Moses truecase.perl program [20].

The output result was text.tru.en and text.true.wl.

iii. Cleaning

Performing the cleaning process Long sentences and empty sentences are eliminated as they can cause problems in the training process, and misaligned sentences are also eliminated by the cleaning [15]. 1 or 50 for the parallel corpus of English-Wolaytta and empty sentences and over 50 words are removed.

The program used to clean corpus using Moses is clean-corpus-n.perl and the output was text.clean.en and text.clean.wl

C. Experimentation

In our modeling, the bilingual English-Wolaytta dictionary uses the hybrid machine translation system; the experiment was carried out in four (4) phases.

Phase 1: Experiment on a Parallel Corpus

In order to develop the hybrid model for the English-Wolaytta languages, we need the parallel corpus for training, tuning, and testing. The prepared parallel corpus must go through various stages before it is used directly for the modeling and training of languages such as tokenization, true casing, and cleaning, so in this phase, we will focus on how these phases were carried out.

D. Corpus Tokenization

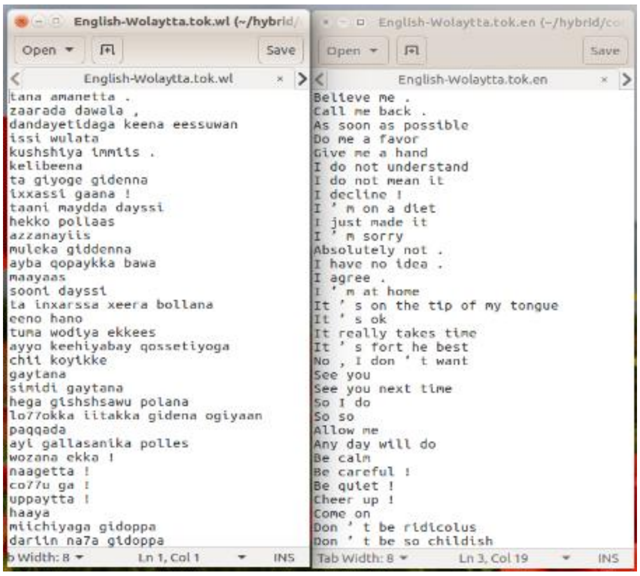
Tokenization was performed both in English and Wolaytta monolingual corpus, thus we put in both monolingual corpora as English-Wolaytta.en and English-Wolaytta.wl. After running the following program, the tokenized monolingual corpus was put in the form English-Wolaytta.tok.en and

Retrieval Number: 100.1/ijsce.B102814020125
DOI: 10.35940/ijsce.B1028.15020525
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-15 Issue-2, May 2025

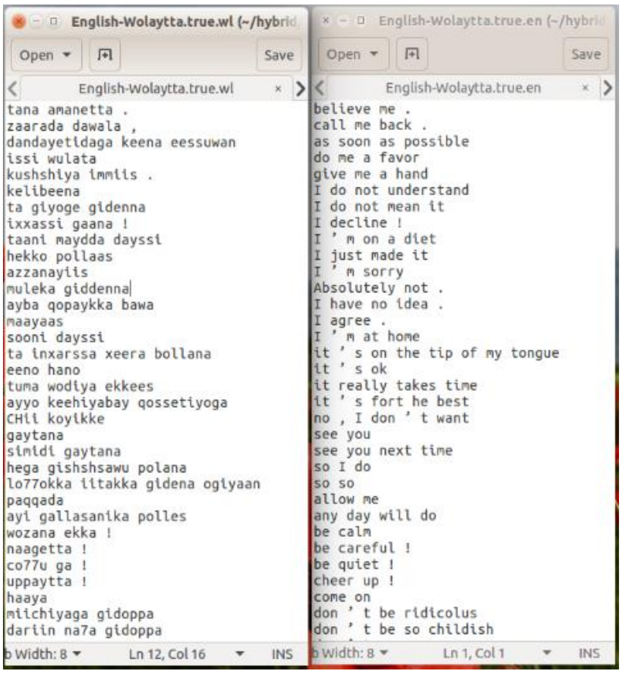
English-Wolaytta.tok.wl. The result of the sample tokenized parallel corpus is shown in Figure 2.



[Fig.2: Sample Tokenized English-Wolaytta Parallel Corpus]

E. Corpus True Casing

The true casing was performed in both English and Wolaytta monolingual corpus, the program has taken both tokenized monolingual corpus as English-Wolaytta.tok.en and English-Wolaytta.tok.wl. The True casing needs a free request program which is a true case modeling in both English monolingual corpus and Wolaytta monolingual corpus. After the program was executed, the true case monolingual corpus was obtained in the form of a true case model. En and true case-model.wl, in the specified working directory. Once the true cased model is executed, the next steps are true casing, normal corpus from the true cased model, and Tokenized one. The sample true-cased parallel corpus is shown in Figure 3.

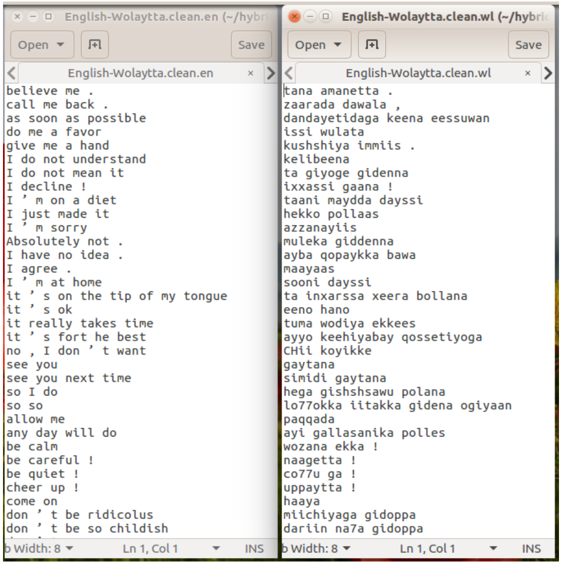


[Fig.3: Sample Cased English-Wolaytta Parallel Corpus]

F. Corpus Cleaning

Corpus cleaning was executed in both English and Wolaytta monolingual corpus. The code used to execute corpus cleaning in Moses scripts is clean-corpus-n.perl.

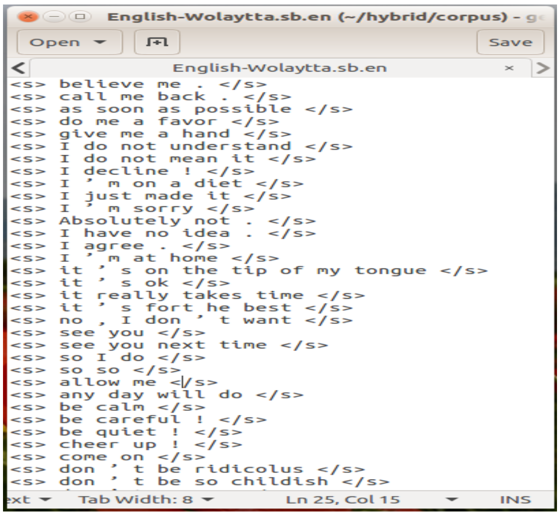
When we executed the cleaning program on our corpus it was obtained in the form of English-Wolaytta.clean.en and English-Wolaytta.clean.wl in a specified working directory.



[Fig.4: Sample Cleaned Parallel Corpus]

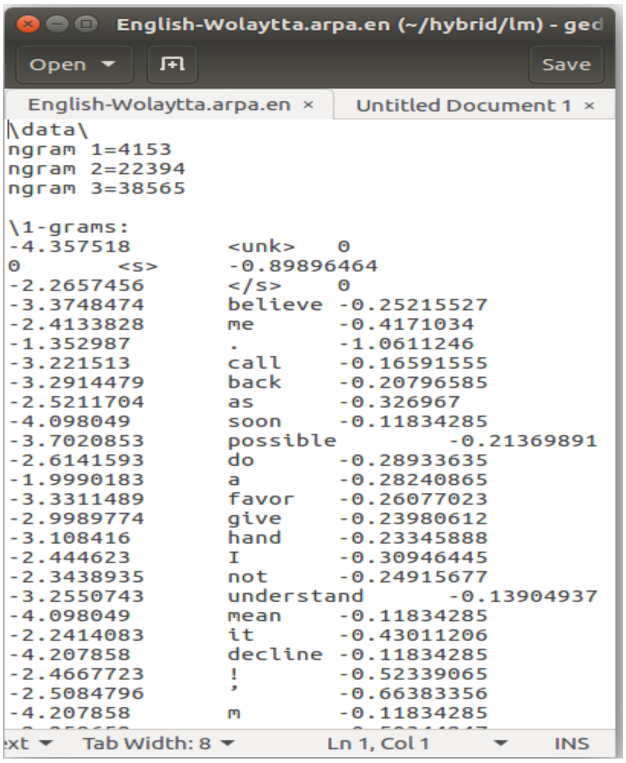
Phase 2: Experiment on Language Modeling

The tool we use to model the language is IRSTLM. It handles LM formats, which makes it possible to reduce both memory and decoding and save time loading. LM Provides tools for building LM, quantizing LM, compiling LM into binary format, accessing binary LM via query class and memory allocation mechanism, and fragment LM [15]. In this phase the **first step** we have flattened the boundary symbol and added the output as English-Wolaytta.sb.en in Figure 5. As the English corpus was smoothed and the border, symbol was added, the same for Wolaytta's text was smooth, and the border symbol was added.



[Fig.5: Sample Smoothed and Boundary Symbol Added Model for English Sentences]

The **second step** was building an appropriate 3-gram language model, it was executed by inputting previously processed data (i.e. English-Wolaytta.sb.en) then we obtained a 3-gram



```
English-Wolaytta.arpa.en (~/hybrid/lm) - ged
English-Wolaytta.arpa.en ×   Untitled Document 1 ×
\data\
ngram 1=4153
ngram 2=22394
ngram 3=38565

\1-grams:
-4.357518       <unk>   0
0       <s>     -0.89896464
-2.2657456      </s>    0
-3.3748474      believe -0.25215527
-2.4133828      me      -0.4171034
-1.352987       .       -1.0611246
-3.221513       call    -0.16591555
-3.2914479      back    -0.20796585
-2.5211704      as      -0.326967
-4.098049       soon    -0.11834285
-3.7020853      possible        -0.21369891
-2.6141593      do      -0.28933635
-1.9990183      a       -0.28240865
-3.3311489      favor   -0.26077023
-2.9989774      give    -0.23980612
-3.108416       hand    -0.23345888
-2.444623       I       -0.30946445
-2.3438935      not     -0.24915677
-3.2550743      understand      -0.13904937
-4.098049       mean    -0.11834285
-2.2414083      it      -0.43011206
-4.207858       decline -0.11834285
-2.4667723      !       -0.52339065
-2.5084796      ,       -0.66383356
-4.207858       m       -0.11834285
xt   Tab Width: 8   Ln 1, Col 1   INS
```

[Fig.6: Sample 3-Gram Language Model]

The third step of language modeling is binarising the 3-gram ARPA language model to faster loading by using KenLM. Once the program is executed, the output binarised model is.blm which is in the form of 0 and 1.

<u>Phase 3: Experiment with Training the Translation Model.</u>

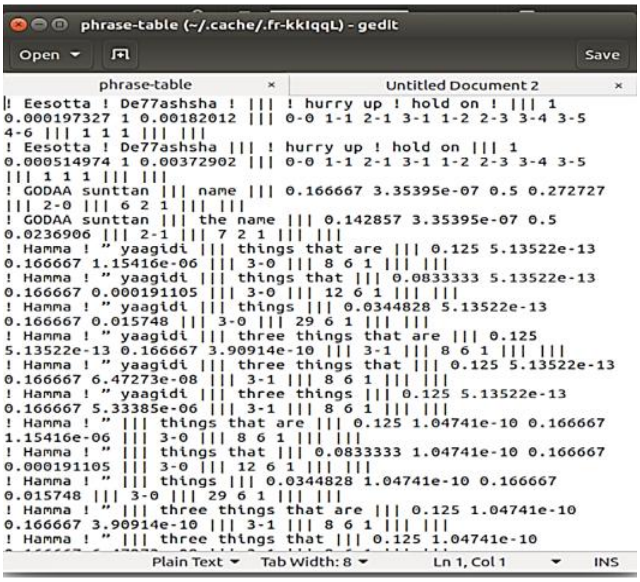
After all, our job is to train the translation model. The Moses Toolkit does an excellent job of completing calls to Giza ++ within a training script and generating the alignment table (phrase table), lexicalized reordering tables, and the Moses configuration file needed for decoding. The Moses decoder begins its execution by going to a directory where Giza ++ is installed.

After running the program there should be a phrase table (alignment table), a lexicalized reordering table, and a configuration file (moses.ini) in the directory home / elsaye / hybrid/working/train/model.

G. Phrase Table

The phrase tables are the main knowledge source for the machine translation decoder Moses.

The decoder consults these tables to figure out how to translate input in the English language into output in the Wolaytta language and vice versa. In the phrase table, each English word is aligned with a possible Wolaytta, and each Wolaytta word is aligned with a possible English word including alignment information. The sample phrase table is shown in Figure 7.

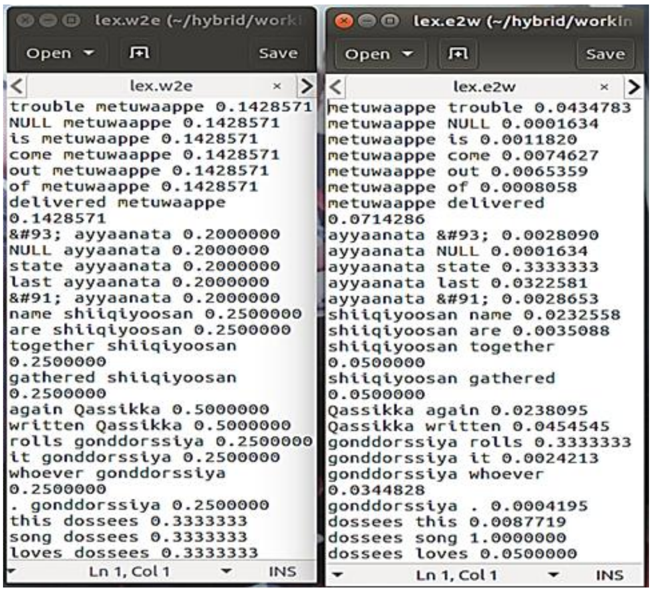


```
phrase-table (~/.cache/.fr-kkIqqL) - gedit
phrase-table ×   Untitled Document 2 ×
! Eesotta ! De77ashsha ! ||| ! hurry up ! hold on ! ||| 1
0.000197327 1 0.00182012 ||| 0-0 1-1 2-1 3-1 1-2 2-3 3-4 3-5
4-6 ||| 1 1 1 ||| |||
! Eesotta ! De77ashsha ||| ! hurry up ! hold on ||| 1
0.000514974 1 0.00372902 ||| 0-0 1-1 2-1 3-1 1-2 2-3 3-4 3-5
||| 1 1 1 ||| |||
! GODAA sunttan ||| name ||| 0.166667 3.35395e-07 0.5 0.272727
||| 2-0 ||| 6 2 1 ||| |||
! GODAA sunttan ||| the name ||| 0.142857 3.35395e-07 0.5
0.0236906 ||| 2-1 ||| 7 2 1 ||| |||
! Hamma ! ” yaagidi ||| things that are ||| 0.125 5.13522e-13
0.166667 1.15416e-06 ||| 3-0 ||| 8 6 1 ||| |||
! Hamma ! ” yaagidi ||| things that ||| 0.0833333 5.13522e-13
0.166667 0.000191105 ||| 3-0 ||| 12 6 1 ||| |||
! Hamma ! ” yaagidi ||| things ||| 0.0344828 5.13522e-13
0.166667 0.015748 ||| 3-0 ||| 29 6 1 ||| |||
! Hamma ! ” yaagidi ||| three things that are ||| 0.125
5.13522e-13 0.166667 3.90914e-10 ||| 3-1 ||| 8 6 1 ||| |||
! Hamma ! ” yaagidi ||| three things that ||| 0.125 5.13522e-13
0.166667 6.47273e-08 ||| 3-1 ||| 8 6 1 ||| |||
! Hamma ! ” yaagidi ||| three things ||| 0.125 5.13522e-13
0.166667 5.33385e-06 ||| 3-1 ||| 8 6 1 ||| |||
! Hamma ! ” ||| things that are ||| 0.125 1.04741e-10 0.166667
1.15416e-06 ||| 3-0 ||| 8 6 1 ||| |||
! Hamma ! ” ||| things that ||| 0.0833333 1.04741e-10 0.166667
0.000191105 ||| 3-0 ||| 12 6 1 ||| |||
! Hamma ! ” ||| things ||| 0.0344828 1.04741e-10 0.166667
0.015748 ||| 3-0 ||| 29 6 1 ||| |||
! Hamma ! ” ||| three things that are ||| 0.125 1.04741e-10
0.166667 3.90914e-10 ||| 3-1 ||| 8 6 1 ||| |||
! Hamma ! ” ||| three things that ||| 0.125 1.04741e-10
Plain Text   Tab Width: 8   Ln 1, Col 1   INS
```

[Fig.7: Sample Phrase Table]

H. Lexicalized Reordering Table

While we perform translation, the lexicalized reordering table has been combined with the phrase table, and Moses configuration file to estimate the correct translation. The probability of each English word given that Wolaytta word is calculated in the lexical reordering table. Figure 8 shows a lexicalized reordering table.



```
lex.w2e (~/hybrid/work
lex.w2e ×
trouble metuwaappe 0.1428571
NULL metuwaappe 0.1428571
is metuwaappe 0.1428571
come metuwaappe 0.1428571
out metuwaappe 0.1428571
of metuwaappe 0.1428571
delivered metuwaappe
0.1428571
&#93; ayyaanata 0.2000000
NULL ayyaanata 0.2000000
state ayyaanata 0.2000000
last ayyaanata 0.2000000
&#91; ayyaanata 0.2000000
name shiiqiyoosan 0.2500000
are shiiqiyoosan 0.2500000
together shiiqiyoosan
0.2500000
gathered shiiqiyoosan
0.2500000
again Qassikka 0.5000000
written Qassikka 0.5000000
rolls gonddorssiya 0.2500000
it gonddorssiya 0.2500000
whoever gonddorssiya
0.2500000
. gonddorssiya 0.2500000
this dossees 0.3333333
song dossees 0.3333333
loves dossees 0.3333333
Ln 1, Col 1   INS
```

```
lex.e2w (~/hybrid/workin
lex.e2w ×
metuwaappe trouble 0.0434783
metuwaappe NULL 0.0001634
metuwaappe is 0.0011820
metuwaappe come 0.0074627
metuwaappe out 0.0065359
metuwaappe of 0.0008058
metuwaappe delivered
0.0714286
ayyaanata &#93; 0.0028090
ayyaanata NULL 0.0001634
ayyaanata state 0.3333333
ayyaanata last 0.0322581
ayyaanata &#91; 0.0028653
shiiqiyoosan name 0.0232558
shiiqiyoosan are 0.0035088
shiiqiyoosan together
0.0500000
shiiqiyoosan gathered
0.0500000
Qassikka again 0.0238095
Qassikka written 0.0454545
gonddorssiya rolls 0.3333333
gonddorssiya it 0.0024213
gonddorssiya whoever
0.0344828
gonddorssiya . 0.0004195
dossees this 0.0087719
dossees song 1.0000000
dossees loves 0.0500000
Ln 1, Col 1   INS
```

[Fig.8: Lexicalized Reordering Table]

I. Moses Configuration File

As a final step, a configuration file for the decoder was generated with all the correct paths in the generated model and a number of default parameter settings, and this file is called moses.ini.

The decoder is controlled by the configuration file moses.ini in addition to this translation model files and language model files are also specified in moses.ini Figure 9. Shows Moses configuration file (moses.ini).

OPEN ACCESS
International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-15 Issue-2, May 2025

moses.ini

```
#########################
### MOSES CONFIG FILE ###
#########################

# input factors
[input-factors]
0

# mapping steps
[mapping]
0 T 0

[distortion-limit]
6

# feature functions
[feature]
UnknownWordPenalty
WordPenalty
PhrasePenalty
PhraseDictionaryMemory name=TranslationModel0 num-features=4 path=/home/elsaye/working/trai
factor=0
LexicalReordering name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allf
working/train/model/reordering-table.wbe-msd-bidirectional-fe.gz
Distortion
KENLM name=LM0 factor=0 path=/home/elsaye/lm/English-Wolaytta.bln.wl order=3

# dense weights for feature functions
[weight]
# The default weights are NOT optimized for translation quality. You MUST tune the weights.
# Documentation for tuning is here: http://www.statmt.org/moses/?n=FactoredTraining.Tuning
UnknownWordPenalty0= 1
WordPenalty0= -1
PhrasePenalty0= 0.2
TranslationModel0= 0.2 0.2 0.2 0.2
LexicalReordering0= 0.3 0.3 0.3 0.3 0.3 0.3
Distortion0= 0.3
LM0= 0.5
```

[Fig.9: Moses Configuration File]

Phase 4: Experiment with tuning the Translation System

Tuning requires a little parallel data, separate from the training data; we have prepared 365 parallel sentences between English and Wolaytta. This corpus has been stored in the home directory as ~/hybrid/corpus/tune/English-tune.en and ~/hybrid/corpus/test/Wolaytta-tune.wl before the tuning process is to be tokenized and true cased. It takes the same procedure as in section 5.3.1 to tokenize and true case tuning parallel corpus.

Once the tokenization and true casing are finished, we make a binarized model for the Phrase table and Reordering table. The following program was used to binarise-model for phrase-table and Reordering-table.

Once the binarized model of phrase-table and Reordering-table is finished, the next step is copying the moses.ini file from Mert-work to binarised-model directory and edit PhraseTableMemory into PhraseTableCompact and change the directory of PhraseTableCompact and LexicalReordering path to Binarised directory, and change the path of both PhraseTableCompact and LexicalReordering to binarised-model directory. Once the program was executed, we went back to the directory used for training, and the tuning process was launched using the Minimum error rate training (MERT) script; it was available in the Moses toolkit.

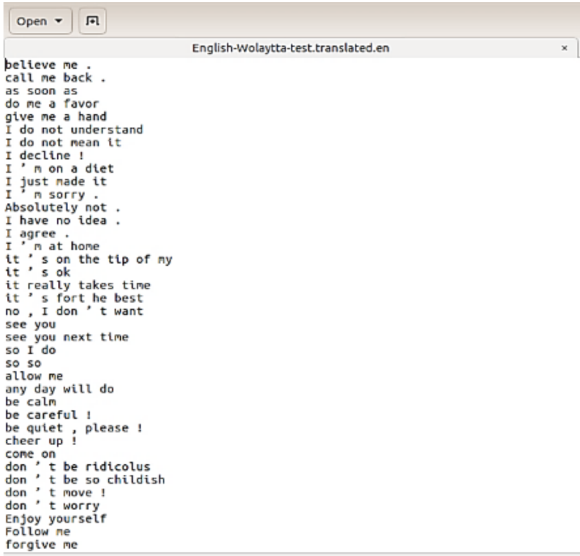
The final result of tuning is a Moses.ini file with trained weights, which should be in ~/working/mert-work/moses.ini. To check the translation with a tuned Moses configuration file, we run the following command line: *~/hybrid$ ~/hybrid/Moses decoder/bin/moses -f ~/hybrid/working/binarised-model/moses.ini* and type in English or Wolaytta sentence, the result of tunes translation displayed on the terminal.

J. Analysis of Translation System

Once the experimentation of the system was finished, analysis of the result that was obtained from the experiments was evaluated. The analysis includes testing and results.

i. Testing the system

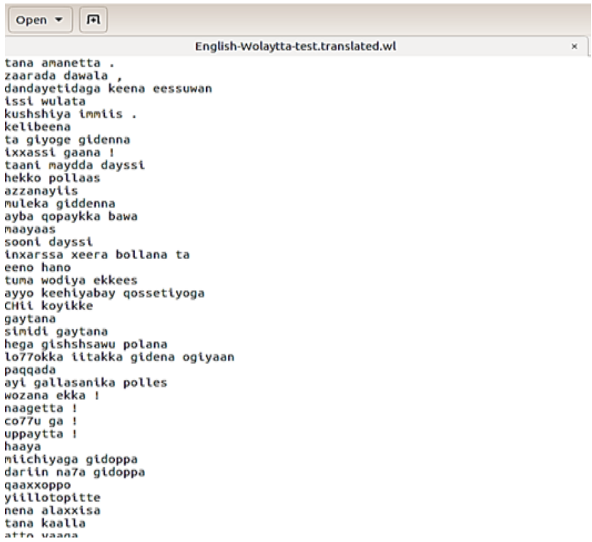
The translation performance evaluation of the Hybrid MT system needs to be measured on a corpus of data different from the training corpus and should include all domains. But, as described in section 3.4 lack of parallel English and Wolaytta data restricts the choice of testing data from all domains, when we tried testing on 100 parallel sentences from the domain other than training and tuning, the obtained translation performance was determined. English-Wolaytta corpus in the domain of training data was created (English-Wolaytta-test.en and English-Wolaytta-test.wl), and it required tokenization and truecasing. This was performed as previously done with the training corpus. Consequently, the Moses script filter-model-given-input.Perl takes English- Wolaytta-test. True.wl testing file to produce the English-Wolaytta-test.translated.en file and the same for the English-Wolaytta-test en and it produces English-Wolaytta-test.wl as shown in Figure 10. a and b.



English-Wolaytta-test.translated.en

```
believe me .
call me back .
as soon as
do me a favor
give me a hand
I do not understand
I do not mean it
I decline !
I ’ m on a diet
I just made it
I ’ m sorry .
Absolutely not .
I have no idea .
I agree .
I ’ m at home
it ’ s on the tip of my
it ’ s ok
it really takes time
it ’ s fort he best
no , I don ’ t want
see you
see you next time
so I do
so so
allow me
any day will do
be calm
be careful !
be quiet , please !
cheer up !
come on
don ’ t be ridicolus
don ’ t be so childish
don ’ t move !
don ’ t worry
Enjoy yourself
Follow me
forgive me
```

[Fig.10a: Translated English Testing File]



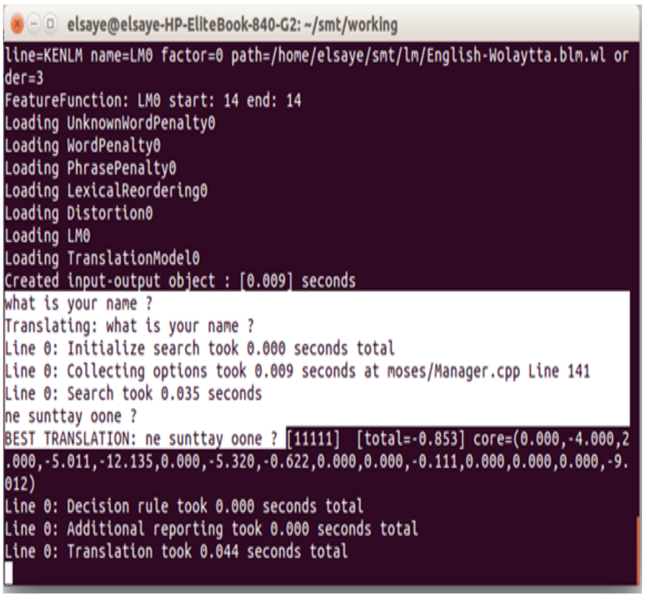
English-Wolaytta-test.translated.wl

```
tana amanetta .
zaarada dawala ,
dandayetidaga keena eessuwan
issi wulata
kushshiya immiis .
kelibeena
ta giyoge gidenna
ixxassi gaana !
taani maydda dayssi
hekko pollaas
azzanayiis
muleka giddenna
ayba qopaykka bawa
maayaas
sooni dayssi
inxarssa xeera bollana ta
eeno hano
tuma wodiya ekkees
ayyo keehiyabay qossetiyoga
CHii koyikke
gaytana
simidi gaytana
hega gishshsawu polana
lo77okka iitakka gidena ogiyaan
paqqada
ayi gallasanika polles
wozana ekka !
naagetta !
co77u ga !
uppaytta !
haaya
miichiyaga gidoppa
dariin na7a gidoppa
qaaxxoppo
yiillotopitte
nena alaxxisa
tana kaalla
```

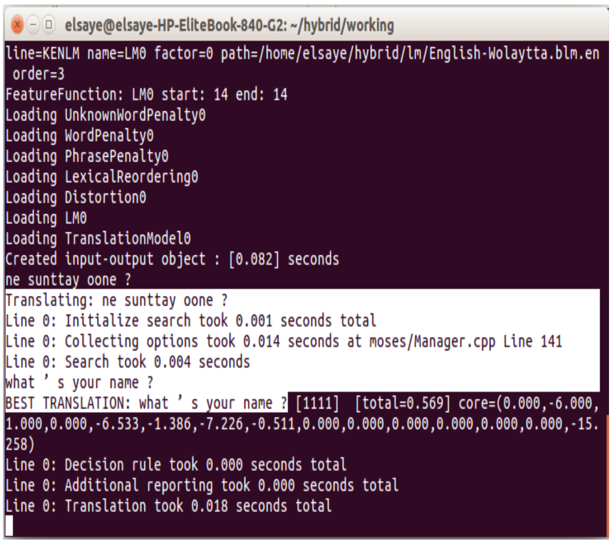
[Fig.10b: Translated Wolaytta Testing File]

The final result of testing is a Moses.ini file with trained weights, which should be In ~/working/filtered-test/moses.ini. To check the translation with a tested

Moses configuration file, we run the following command line:
~/hybrid/Moses decoder/bin/moses -f ~/hybrid/working/filtered-test/moses.ini and type in English or Wolaytta sentence, the result of testing translation displayed on the terminal as shown in Figure 11 a, and 11 b



[Fig.11a: Sample of English to Wolaytta Translated Sentence]



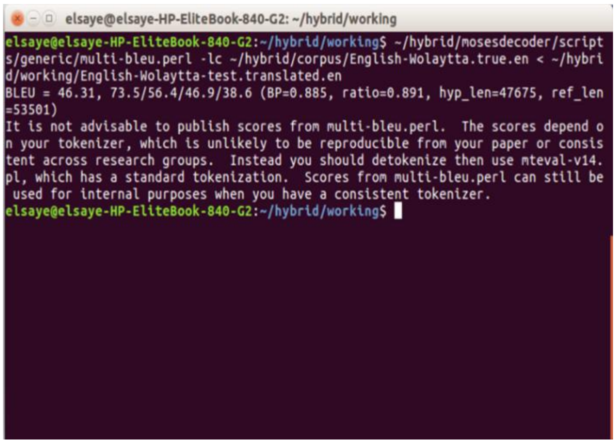
[Fig.11b: Sample of Wolaytta into English Translated Sentence]

ii.Results

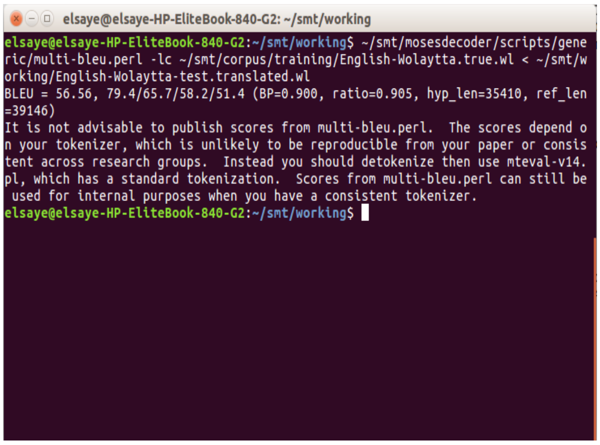
The English-Wolaytta hybrid MT system was trained and evaluated using the English-Wolaytta parallel training set as translation examples and the English and Wolaytta Source Text as new sentences, yielding an output Target Text comprising translated English and Wolaytta sentences (Figure 11a and 11b).

The BLEU evaluation measures were used to rate the output. Testing the system has been performed in two states: which is from English into Wolaytta and from Wolaytta into English.

The comparisons of the experimental results are translating from Wolaytta to English and translating from English to Wolaytta is discussed below.



[Fig.12a: Bleu Score from Wolaytta English Sentence]



[Fig.12b: Bleu Score from English to Wolaytta Sentence]

Table 1: Detailed Summary of the English-Wolaytta Machine Translation Experiment

Language	Prepared Sentences				Bleu Score	
	For Training	For Tuning	For Testing	Total	English to Wolaytta	Wolaytta to English
English	3004	365	100	3,469	56.56%	46.31%
Wolaytta	3004	365	100	3,469		

Table 2: Few Examples of English Wolaytta Translations

English	Wolaytta
Believe me.	tana amanetta .
Call me back.	zaarada dawala ,
as soon as possible	dandayetidaga keena eessuwan
do me a favor	issi wulata
give me a hand	kushshiya immiis .
I do not understand	Gelibeena

V. CONCLUSION

In this paper, we developed a Translation system for Bidirectional English-Wolaytta machine translation using a hybrid approach. A system is made up of four parts: sentence reordering, language modeling, decoding, and translation modeling. By applying its POS tagging,

Retrieval Number: 100.1/ijsce.B102814020125
DOI: [10.35940/ijsce.B1028.15020525](https://doi.org/10.35940/ijsce.B1028.15020525)
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

preprocess the source language's structure so that it is more comparable to the structure of the destination language and is better handled by the statistics engine. For Wolaytta sentences we manually tagged as there are no easily accessible tag sets. To create reordering rules for different sorts of Wolaytta phrases and sentences in English, the linguistic background and nature of the two languages were researched.

We used the language modeling toolkit IRSTLM which is used to understand the source and target languages. In this study, the Moses decoder was used to decode a system. It takes the source sentence and decodes it into the best sentence of the target language. For the translation model, we use the Giza ++ Toolkit, which is available on the GitHub website. Finally, four phases of experiments are performed on the collected data set to verify the translation of the system. Phase 1: is an experiment in the parallel corpus, phase 2: is an experiment in language modeling, phase 3 is: an experiment in training the translation model, and Phase 4: is an experiment in tune and testing the translation system and we receive both the English to Wolaytta and Wolaytta to English translation.

For both lateral translations, the Bleu score assessed the accuracy of the translation. Wolaytta to English was 46.31% and from English to Wolaytta was 56.56%. Driven by the result from English to Wolaytta, they have better accuracy than the Wolaytta to English language pair because for English words there is more than one translation of Wolaytta word.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Kim, M. K., Takero, H., & Fedovik, S. (2023). Universal Syntactic Structures: Modeling Syntax for Various Natural Languages. arXiv preprint arXiv:2402.01641. DOI: https://doi.org/10.48550/arXiv.2402.01641
2. Ashkanasy, N. M., Trevor-Roberts, E., & Earnshaw, L. (2002). The Anglo cluster: Legacy of the British empire. Journal of World Business, 37(1), 28-39. DOI: https://doi.org/10.1016/S1090-9516(01)00072-4
3. Bade, G. Y., & Seid, H. (2018). Development of Longest-Match Based Stemmer for Texts of Wolaita Language. Vol, 4, 79-83. DOI: https://doi.org/10.11648/j.ijdst.20180403.11
4. Bedecho, A. T., & Bokka, R. K. (2024). Development of Sentiment Analysis for the Wolaita Language using Machine Learning Approaches. In 2024 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 178-182). IEEE. DOI: https://doi.org/10.1109/ICT4DA62874.2024.10777118
5. Ambushe, S. A., Awoke, N., Demissie, B. W., & Tekalign, T. (2023). Holistic nursing care practice and associated factors among nurses in public hospitals of Wolaita zone, South Ethiopia. BMC nursing, 22(1), 390. DOI: https://doi.org/10.1186/s12912-023-01517-0
6. Rossi, C. (2017). Introducing statistical machine translation in translator training: from uses and perceptions to course design, and back again. Revista Tradumàtica: tecnologies de la traducció, (15), 48. Doi : https://doi.org/10.5565/rev/tradumatica.195
7. Azath, M., & Kiros, T. (2020). Statistical machine translator for English to Tigrigna translation. International Journal of Scientific and Technology Research, 9(1), 2095-2099. DOI: https://www.readkong.com/page/statistical-machine-translator-for-english-to-tigrigna-1868057
8. Teshome, E. (2013). Bidirectional English-Amharic machine translation: an experiment using constrained corpus (Doctoral dissertation, Addis Ababa University). DOI: http://thesisbank.jhia.ac.ke/id/eprint/6064
9. Mara, M. (2018). English-Wolaytta Machine Translation using Statistical Approach (Doctoral dissertation, St. Mary's University). http://www.repository.smuc.edu.et/handle/123456789/4462
10. Tulu, G. (2022). Bidirectional AmharicAfaan *Oromo Machine Translation Using Hybrid Approach*. DOI: https://projectng.com/topic/co22921/bidirectional-amharic-afaan-oromo-machine#google_vignette
11. Shirko, B. F. (2020). Part of speech tagging for wolaita language using transformation-based learning (tbl) approach. DOI: https://www.researchgate.net/publication/345243262_Part_of_Speech_Tagging_for_Wolaita_Language_using_Transformation_based_Learning_TBL_Approach
12. Sinhal, R. A., & Gupta, K. O. (2014). Machine translation approaches and design aspects. IOSR Journal of Computer Engineering, 16(1), 22-25. DOI: https://doi.org/10.9790/0661-16122225
13. Koehn, P. (2009). Statistical machine translation. Cambridge University Press. Doi: https://doi.org/10.1017/CBO9780511815829
14. Chéragui, M. A. (2012). Theoretical Overview of Machine Translation. ICWIT, 160-169. DOI: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=aad01b2a642711ef0b4d7d89d8d50fc268a222ce
15. Phan, H., & Jannesari, A. (2020). Statistical machine translation outperforms neural machine translation in software engineering: why and how Proceedings of the 1st ACM SIGSOFT International Workshop on Representation Learning for Software Engineering and Program Languages, Virtual, USA. DOI: https://doi.org/10.1145/3416506.3423576
16. Thendral, R., & Sigappi, AN. (2020). Stacked Bidirectional Long Short Term Memory Models To Predict Protein Secondary Structure. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 3, pp. 1605–1608). DOI: https://doi.org/10.35940/ijitee.c8368.019320
17. Vidya, K., Annapoorani, P., Akila, S., & Vijayalakshmi, M. (2019). Microcontroller Based Bi-Directional DC-DC Converter for Automobile Application. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 2, pp. 2776–2778). DOI: https://doi.org/10.35940/ijeat.b2280.129219
18. S. T. Shenbagavalli, D. Shanthi, S. Naganandhini, R. Karthikeyan, Role of Deep Recurrent Neural Networks in Natural Language Processing. (2019). In International Journal of Recent Technology and Engineering (Vol. 8, Issue 2S11, pp. 4082–4084). DOI: https://doi.org/10.35940/ijrte.b1597.0982s1119
19. Krishna, G. G. (2023). Multilingual NLP. In International Journal of Advanced Engineering and Nano Technology (Vol. 10, Issue 6, pp. 9–12). DOI: https://doi.org/10.35940/ijaent.e4119.0610623
20. Patidar, C. P., Katara, Y., & Sharma, Dr. M. (2020). Hybrid News Recommendation System using TF-IDF and Similarity Weight Index. In International Journal of Soft Computing and Engineering (Vol. 10, Issue 3, pp. 5–9). DOI: https://doi.org/10.35940/ijsce.c3471.1110320

AUTHORS PROFILE

Elisaye Bekele Milke, earned a B.Sc. in Information Technology from Wolaita Sodo University, Ethiopia, in 2016, and a M.Sc. from Ethiopian Technical University in 2021. In 2021, he joined Wolaita Sodo University as a Lecturer in the Department of Information Technology. His research focuses on Natural Language Processing, Deep Learning, Machine Learning, Data

Mining, image processing, and Web Technology. He is dedicated to advancing knowledge and innovation in these fields, contributing to both academic excellence and practical solutions.

Tibebe Beshah Tesema (PhD) is an Associate Professor of Information Systems and Head of the School of IS at Addis Ababa University, Ethiopia. He also coordinates the IS track of the IT Doctoral Program. His research focuses on data, web, and sentiment mining, information architecture, knowledge representation, and the appropriation of information systems in organizations. He has authored or co-authored over 55 scientific articles and led numerous research projects in information systems.

Mesfin Leranso Betalo (PhD), earned a B.Sc. in Computer Science from Hawassa University, Ethiopia, in 2012, and an M.Sc. in Information Technology from Madras University, India, in 2016. He completed a Ph.D. in Information and Communication Engineering at UESTC, China. From 2018–2022, he was a lecturer and postgraduate committee member at Wachemo University, Ethiopia. He joined the Ubiquitous and Wireless Networks research team at UESTC and is now a postdoctoral fellow at Shenzhen University, China. His research focuses on the Internet of Vehicles, 6G networks, IoT, intelligent transportation, autonomous aerial vehicles, machine learning, cloud computing, and mobile robots.

Retrieval Number: 100.1/ijsce.B102814020125
DOI: 10.35940/ijsce.B1028.15020525
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)