

Enhancing Bi-directional English-Tigrigna Machine Translation Using Hybrid Approach

Zemicheal Berihu¹, Gebremariam Mesfin^{1,2}, Mulugeta Atsibaha¹, Tor-Morten
Grønli²

¹Aksum University, Department of Computing Technology, Aksum, Ethiopia

²Kristiania University College, Department of Technology, Oslo, Norway

Abstract

Machine Translation (MT) is an application area of NLP where automatic systems are used to translate text or speech from one language to another while preserving the meaning of the source language. Although there exists a large volume of literature in automatic machine translation of documents in many languages, the translation between English and Tigrigna is less explored. Therefore, we proposed the hybrid approach to address the challenges of applying syntactic reordering rules which align and capture the structural arrangement of words in the source sentence to become more like the target sentences. Two language models were developed- one for English and another for Tigrigna and about 12,000 parallel sentences in four domains and 32,000 bilingual dictionaries were collected for our experiment. The parallel collected corpus was split randomly to 10,800 sentences for training set and 1,200 sentences for testing. Moses open source statistical machine translation system has been used for the experiment to train, tune and decode. The parallel corpus was aligned using the Giza++ toolkit and SRILM was used for building the language model. Three main experiments were conducted using statistical approach, hybrid approach and post-processing technique. According to our experimental result showed good translation output as high as 32.64 BLEU points Google translator and the hybrid approach was found most promising for English-Tigrigna bi-directional translation. **Keywords:** Machine translation, English, Tigrinya, Bilingual dictionary, BLEU, Corpus, HMT, Syntactic re-ordering rule

Introduction

Natural Language Processing (NLP) is an interdisciplinary field that discovers how computers can be used to recognize and operate natural language text or speech to do valuable things[1]. Even though NLP is considered as a growing and hot research area with its own accumulative history since 1950s[2], it has mostly continued within computational linguistics domain. The intention behind NLP based research is to make computer systems, to recognize and operate natural languages to perform the anticipated tasks. It is pertinent in numerous domains; such as Cross Languages Information Retrieval (CLIR), Machine Translation (MT), text processing, text summarization, speech recognition, user interfaces, and artificial intelligence. MT, the most and hot research area of NLP, where automatic systems are used to translate text from one language to another while maintaining the meaning of the source language. According to [3], it is an amalgamation of areas such as computational linguistics, artificial intelligence, translation theory and statistics. It is also fast and easy to use whereas human translation is very slow, cumbersome, time consuming and expensive task as compared to MT.

MT applications can be categorized according to their core methodology[4]. The most commonly used one is known as the rule-based, and the statistical approach. In the rule-based approach, human experts stipulate a set of rules to describe the translation process, so that a huge amount of input from human experts is required. In contrast, under the statistical approach the knowledge is automatically mined by analyzing translation samples from a parallel corpus built by human experts. Merging the features of the two major categories of MT systems gave birth to the hybrid approach. Hence, in this research work, the researcher come up with a solution of integrating rule based followed by statistical approach were conducted as a sequence of rule based followed by the statistical approach so as to accommodate the advantage of both approaches.

Overview of the Tigrinya language

Tigrigna is one of the low-resource languages and belongs to the Semitic language family of the Afro Asiatic phylum, along with Hebrew, Amharic, Maltese, Tigre, and Arabic. Tigrigna claims an

estimated 7 million speakers in Eritrea and northern Ethiopia. Unlike major Semitic languages, which enjoyed relatively widespread NLP research and resources, Tigrigna was largely ignored in NLP-related research due to the absence of a readymade corpus [5]. It is a widely spoken language in Eritrea and in the northern part of Ethiopia. In Eritrea it is a working language in offices along with Arabic. Tigrigna is also spoken by many immigrant communities around the world, such as in the Sudan, Saudi Arabia, the United States, Germany, Italy, United Kingdom, Canada, and Australia. Tigrigna is written in the Ge'ez script, originally developed for the now-extinct Ge'ez. There are 32 set of letters in Tigrigna alphabet (“ATLAS - Tigrigna”, n.d.).



Figure 1 Countries in which Tigrigna is spoken (“ATLAS - Tigrigna”, n.d.)

The Tigrinya writing system

The Tigrigna writing system is one variant of what is often referred to as the "Ethiopic" writing system or "Ethiopic syllabary". It is a slight variant of the writing system used for Tigrigna and for Ge'ez, the classical language still in use as the liturgical language of Ethiopian and Eritrean Orthodox Christians (“The Tigrigna Writing System,” n.d.). Generally, Tigrigna writing system is like English writing system from left to right. In this section the construction of Tigrigna sentence and types of sentences are described. Unlike English, the sentence structure for Tigrigna is a Subject-Object-Verb (SOV) whereas English is Subject-Verb-Object (SVO) amalgamation.

Research methodology

This research is based on quantitative experimental method which involves corpus preparation, tool selection for building language model, translation model and evaluation of the performance of the model.

The Hybrid Approach to Machine Translation (HMT) maximizes the strength of both the statistical and rule-based approaches. Numerous MT corporations (Asia Online and Systran) are claiming to have a hybrid approach. The methodologies vary in some various methods. Rules are post-processed by statistical approach that is translations are accomplished using rules and then the statistics are applied to adjust the output from the rule's engine. Statistics are then guided by rules that means different rules are then helps to pre-process the corpus in try to well guide the statistical approach. Several rules are used then to post-process the statistical output to accomplish pre-processing steps such as normalization. This approach has a lot more power, flexibility and control when translating[9]. The main motivation behind using of hybrid approach for this research work is to escape from the failure of any single system to achieve satisfactory results. Because hybrid solutions can tend to combine the advantages of individual approaches to achieve an overall better translation[10].

Architecture of the proposed system

The proposed system is divided into modules, each of which performs different tasks, where English and Tigrigna texts are locally reordered before applying statistical methods for translation. The proposed Pipeline based architecture used to manage the flow of information across the modules is given in Figure 2 below.

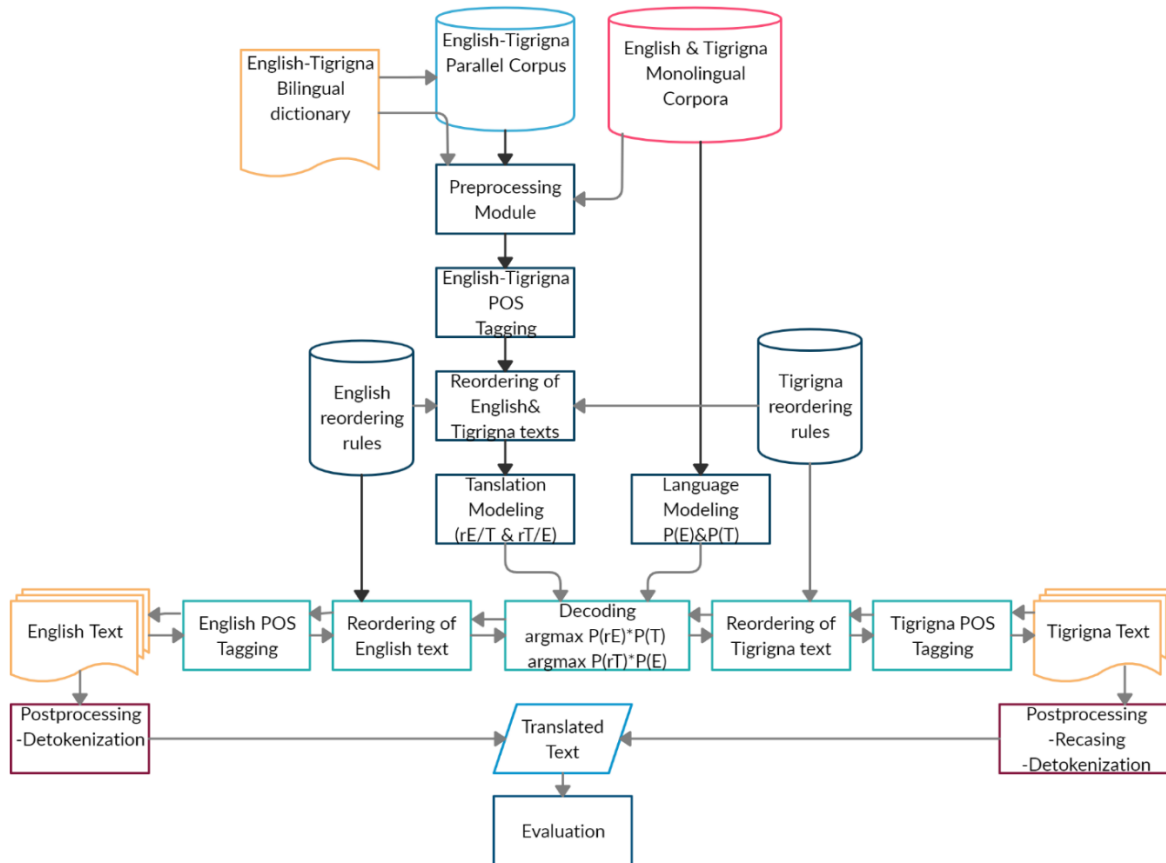


Figure 2 Bi-directional English-Tigrigna MT System Architecture

We use English POS tagger and existing Tigrigna tagger to identify English and Tigrigna word class marker respectively, which help to apply local reordering rules. After the English text is locally reordered takes the form of Tigrigna text and after the Tigrigna text is locally reordered, it takes the form of English text, then after we applied the statistical methods for translation. The translation model is built using the parallel corpus of Tigrigna text and locally reordered English text and is built using the parallel corpus of English and locally reordered Tigrigna texts. The language model is built using corpus of both English and Tigrigna text since the translation is two-way directions which is from English-to-Tigrigna and vice versa.

Corpus preprocessing

Corpus pre-processing is a process of readying and compiling of data for experimentation processing. The parallel corpus even though is published by Bible book, Constitution, Tourism and News publishers, to better recognize and understand by the system being developed the need to be cleaned is non-questionable. In this stage, blank lines, sentences that have no equivalent translation are removed from the parallel corpus.

Bilingual sentence alignment

Sentence alignment is crucial in machine translation and it is too hard problem, but in this research, it is simple by the concept that most of the texts are already available in sentence and verse aligned format. The sentence level aligned corpus is stored in one file per language, so that lines with the same line number in a file pair are mappings of each other. In this research, the size of the sentence level aligned corpus is 12,000 sentences per language.

Corpus tokenization

Tokenization is one of the normalization steps conducted in this research work, which insert spaces between words and punctuations, splits longer strings of English and Tigrigna text into smaller pieces,

or tokens. It is also referred to as text segmentation or lexical analysis. Occasionally segmentation is used to refer to the breakdown of a large chunk of text into a number of pieces larger than words (paragraphs or sentences), whereas tokenization is held in reserve for the breakdown process which results exclusively in words. Further text processing is generally accomplished after a piece of text has been appropriately tokenized.

Corpus normalization

Normalization is a pre-processing step which denotes to a series of related tasks meant to put all text to the same case, removing punctuation, and converting numbers to their word equivalents. In this research the Ethiopic script includes different letters that have the same sound in Tigrigna language. According to [11], the letters 'ኅ'(he) and 'ሀ' (he),'ሰ' (se) and 'ሠ' (se), letters 'ጸ' (Tse) and ፀ (Tse) are some examples. Although most Tigrinya writings have one of these forms, some writers use them interchangeably. In Eritrea the 'ጸ' series is used while in Ethiopia the ፀ series is used. Thus, a single Tigrinya word may exist in two different variations on many web documents. For example, ጠፀጵኔት (meShEt) and ጠፀፀኔት (me'ShEt) are two variants of the same word meaning 'pamphlet'. Such variant forms have negative effect on precision of retrieval. Thus, normalization process is conducted on the parallel corpora for the sake of overcoming writing inconsistency and enhancing translation quality.

Table 1 Tigrigna Character Normalization

Tigrigna Characters						
ሠ/ሰ	ሠ-/ሰ-	ሠ/ሰ	ሠ/ሰ	ሠ/ሰ	ሠ/ሰ	ሠ/ሰ
ጸ/ፀ	ጸ-/ፀ-	ጸ/ፀ	ጸ/ፀ	ጸ/ፀ	ጸ/ፀ	ጸ/ፀ
ኅ/ሀ	ኅ-/ሀ-	ኅ/ሀ	ኅ/ሀ	ኅ/ሀ	ኅ/ሀ	ኅ/ሀ

As shown from Table 1 the normalization script is converted to the same character (format) for the sake of having consistency in the translation pipeline regardless of the corpus type and magnitude. For instance, ጸልማት and ፀልማት are different in meaning but have the same pronunciation. Therefore, the character “ጸ” should have to change to the character “ፀ” for consistency purpose in the translation pipeline. As a result, there is no confusion during training of the translation system(language model) for both source and target languages due to the successful execution of character level normalization as shown from the above interchangeable usage of characters in Tigrigna writing is shown in Figure3 below.

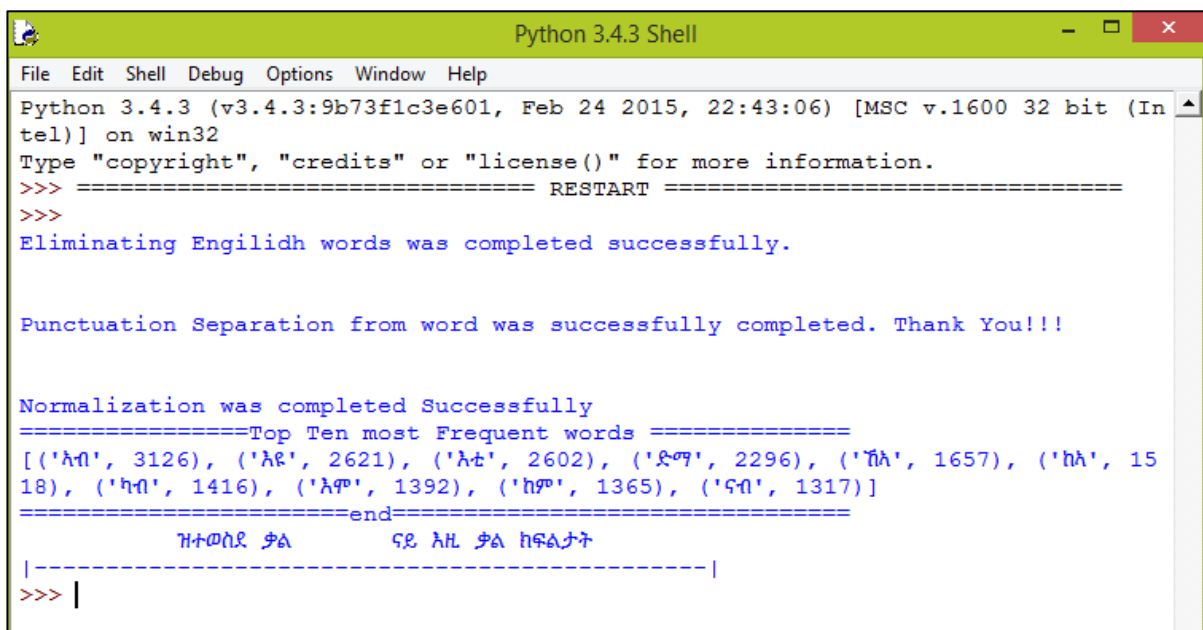


Figure 3 Tigrigna Corpus Normalization Process

Stemming for Tigrinya

This is the other pre-processing step conducted to obtain the stem of those words given in the prepared corpus which are not found in vocabulary and to reduce inflected word forms into common basic form and it is significant analysis procedure in information retrieval and many NLP applications. Not only this but also it is a normalization step that diminishes the morphological forms of words to a common form usually called a stem by the removal of affixes[11]. If stemmed word is present in vocabulary, then that is an actual word, otherwise it may be a proper name or some invalid word. For instance, when a user prompts to enter an input word ኣይሰማዕኹን and if the input word is not present in the vocabulary of the corpus then it may cause erroneous result. Generally, with the assistance of a stemmer, one can reduce the desired word into its root or stem word. In this example, ሰማዕ is the stem or root word ኣይ- is prefix and -ኹን is the suffix. Stem supplies the main meaning of the word while the suffixes add additional meanings. In this study we have used an existing Tigrigna stemmer developed by [12] for Tigrigna corpus to facilitate the information retrieval and to handle out of vocabulary during translation process.

Postprocessing

In this study, post processing is conducted to correct MT output that has not been pre-processed, for instance in order to improve the grammaticality [13]. It changes the translated output from SMT system into standard target language sentence. Basically, post processing is needed after the translation when the target language has been pre-processed, in order to restore it to the normal target language. It can use also on standard MT output, in order to correct some of the errors from the MT system. We have conducted post processing in order to correct and identify in the MT output, both in order to evaluate and compare systems. These post processing techniques are used during this research work were recasing and detokenizing the MT output.

Corpus preparation

In this study to implement the experiments, parallel sentences and bilingual dictionaries were collected from four domains (Bible, Constitution Proclamation, Tourism and News). The reason behind to select these domains of corpus for corpus preparation is, because, the data is easily accessible from the web and they are parallel corpus which is suitable for the SMT pipeline.

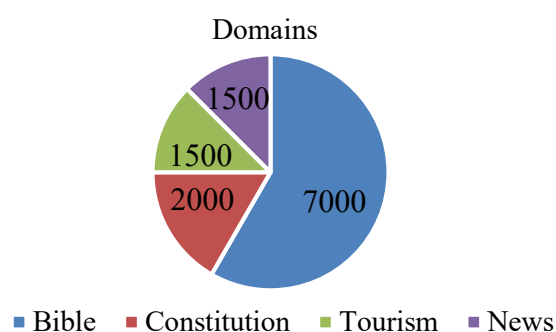


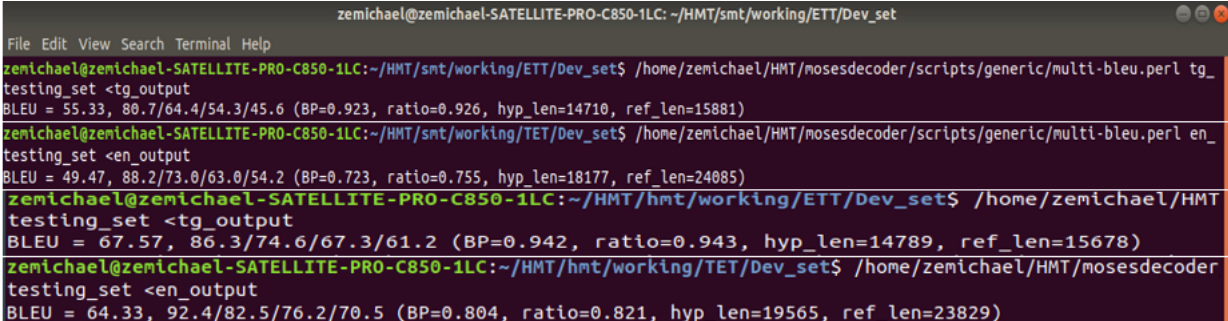
Figure 4 Application domains of sources of the corpus

We executed corpus tokenization, truecasing, cleaning, and normalization during corpus pre-processing stage to make the corpus suitable in format and ready for experimentation. In this study, the size of the corpus used for the experiments is 12,000 parallel sentences and 32,000 bilingual dictionaries 6400 sentences, prepared from the four domains. Out of 12,000 parallel sentences 90% (10,800) and 10% (1200) were selected randomly for training and testing respectively.

Experimental results

The purpose of this experiment is to train the translation system to create a model depends on the trained corpus. Three major experiments were conducted. The first experiment is accompanied on English-to-Tigrigna translation by applying a statistical approach. The second experiment accompanied on English-to-Tigrigna translation by applying a post-processing approach; and the third experiment is accompanied on English-to-Tigrigna translation by applying the hybrid approach. After testing process is accomplished, the test results for the three major experiments were recorded and discussed as follow:

For all experiments, in order to measure the performance of the trained moses.ini system, the model must be tested using gold dataset outside of the train dataset. Since test datasets are not part of the training dataset, out of 12,000 English and Tigrigna parallel sentences 1,200 are selected randomly from the total corpus using the m4loc testset.pl script depend on [14] and [15] in order to test the performance of the system in terms of translation accuracy to translate a single English sentence to Tigrigna sentence. To do so, BLEU score methodology which has been used in order to measure the result of the translation process before and after tuning. Therefore, the recorded result for each experiment is given in the following figures. The first experiment test result is given in Figure 5 and Figure 6 for both translation directions. Accordingly, the result scored from the BLEU score methodology shows 55.33 of the translation is correctly performed from English to Tigrigna texts and 49.47 from Tigrigna to English texts as shown in Figure 5- (a) and (b) respectively.



```
zemichael@zemichael-SATELLITE-PRO-C850-1LC: ~/HMT/smt/working/ETT/Dev_set
File Edit View Search Terminal Help
a. zemichael@zemichael-SATELLITE-PRO-C850-1LC:~/HMT/smt/working/ETT/Dev_set$ /home/zemichael/HMT/mosesdecoder/scripts/generic/multi-bleu.perl tg_
testing_set <tg_output
BLEU = 55.33, 80.7/64.4/54.3/45.6 (BP=0.923, ratio=0.926, hyp_len=14710, ref_len=15881)
b. zemichael@zemichael-SATELLITE-PRO-C850-1LC:~/HMT/smt/working/TET/Dev_set$ /home/zemichael/HMT/mosesdecoder/scripts/generic/multi-bleu.perl en_
testing_set <en_output
BLEU = 49.47, 88.2/73.0/63.0/54.2 (BP=0.723, ratio=0.755, hyp_len=18177, ref_len=24085)
c. zemichael@zemichael-SATELLITE-PRO-C850-1LC:~/HMT/hmt/working/ETT/Dev_set$ /home/zemichael/HMT
testing_set <tg_output
BLEU = 67.57, 86.3/74.6/67.3/61.2 (BP=0.942, ratio=0.943, hyp_len=14789, ref_len=15678)
d. zemichael@zemichael-SATELLITE-PRO-C850-1LC:~/HMT/hmt/working/TET/Dev_set$ /home/zemichael/HMT/mosesdecoder
testing_set <en_output
BLEU = 64.33, 92.4/82.5/76.2/70.5 (BP=0.804, ratio=0.821, hyp_len=19565, ref_len=23829)
```

Figure 5 BLEU Score for English-Tigrigna Bi-directional Translation after Tuning

The result recorded from the BLEU score methodology shows 67.57 of the translation is correctly performed from English to Tigrigna texts and 64.33 from Tigrigna to English texts as shown in Figure 5- (c) and (d) respectively.

Prototype of the system

All the above experiments are conducted on Ubuntu 18.04 terminal that is challenging to use by the end user of the system who have not experience in Ubuntu. In addition, this section will try to demonstrate the prototype of the system when it is queried to translate from English to Tigrigna and vice versa. To run the translation system, the end user needs a knowledge in Ubuntu terminal execution process. But this prototype of translation system uses for everyone easily without any requirement on terminal of the Ubuntu operating system. The solution for the above problem is changing the terminal executable translation process into web-based translation system. In order to perform the webpage, the following software are important. Apache server is used to communicate the terminal executable file with the PHP, in order to design of the webpage, write to file and open from file. The graphical user interface of the translation system shown below.

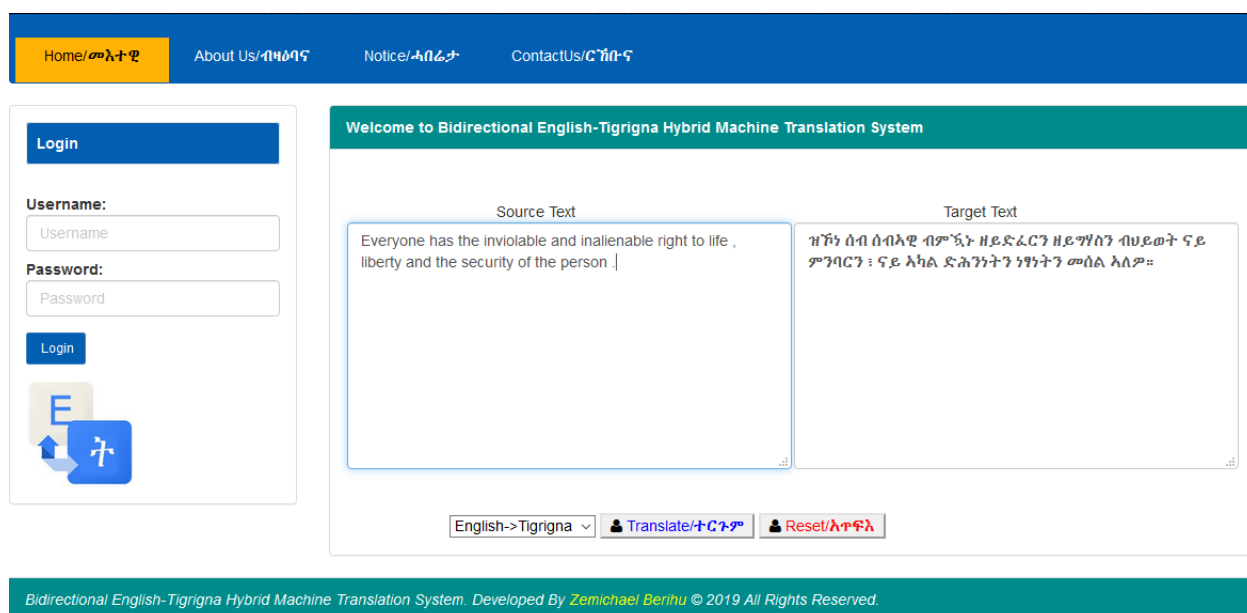


Figure 6 Sample translation using the prototype

The second step is creating shell file that includes the path of the terminal executable. We create com.sh file and the file include the path to run the moses.ini. in this stage in addition to the shell file, two files are need the first one uses to write the text written in the source form and the second file is also to hold the translated text of the input. The third step is included the PHP file into the apache server. As a result, execute the server and insert input text in the source form and when we click the ተርጉም (translate) button the input text would be write into the file created before to uses as input and using the shell file it changes the input text into target, using the model and saved in the second file. At the end, the second file would be open into the target form.

The execution of the webpage shown as Figure 6 below shows that the translation of English sentence “Everyone has the inviolable and inalienable right to life, liberty and the security of the person.” to Tigrigna sentence “ዝኾነ ሰብ ሰብኣዊ ብምኻኑ ዘይድፈርን ዘይግሃስን ብህይወት ናይ ምንባርን ፣ ናይ ኣካል ድሕንነትን ነፃነትን መሰል ኣለዎ።” which is the translation from English to Tigrigna direction using a representative(testing) set of the corpus. Furthermore, this knowledge is extracted from Ubuntu terminal using PHP script and shell extension from moses decoder of the training model located in path, Home/Zemichael/HMT/Working/ETT/Train/Model/moses.ini working directory.

Discussion

Three experiments are conducted with two different approaches and technique. As it is shown in from the above three experiments namely; experiment I, experiment II, and experiment III, one can observe that, the result recorded from BLEU score shows the hybrid approach is better than the statistical approach and post processing technique for both translation directions. Post processing technique made an improvement over statistical approach in the experiment. All experiments are guided by statistical approach and since statistical approach is based on bilingual corpus, as the size of corpus become increase, the accuracy of the system also increase and similarly the BLEU score result can also increase with three sets of randomly selected and observed sentences. In machine translation the performance of the system highly depends on the domain of training and testing set, if the testing set is within the training domain the performance of the system will increase, that means using a lot of training data is useful, if we’ve never seen the phrase “world vision” before in our parallel corpus, then probably the system is not going to translate it correctly.

The BLEU score result of 51.85 and 46.11 is for English-Tigrigna and Tigrigna-English translations respectively, before tuning indicates the testing corpus used is not the same with training set but related. The domain variety of the testing corpus has affected the BLEU score significantly. As most of our corpus is from the holy bible, upon evaluating the output, we have investigated the bias

introduced by the training corpus from this domain. That is, the system performs better if it is tested on constitution corpus than from other available domain for both directions.

The researcher has seen separately, the BLEU score for the test corpus from the news, and tourism domains which have small amount of corpus in training was 1350(12%). But when we prepare testing corpus from all related domain of training dataset, we have achieved the BLEU score of 67.57% and 64.33% for English-Tigrigna and Tigrigna-English. The hybrid system displays definite gains over others. Generally, the result recorded from BLEU Score was 67.57% for the English-Tigrigna translation and 64.33% for the Tigrigna-English translation. The result recorded from English to Tigrigna was low mostly because it was hard for the system to identify the feminine and masculine representation not only this but also the transliteration problem of the two language pairs. From the above table we can conclude that applying post processing techniques and hybridization enhanced the system performance for both directions from the baseline system by 0.72 for English to Tigrigna and 0.74 for Tigrigna to English as well as 12.24 and 14.86 for the same direction respectively. Finally, if the training set and testing set are separated manually, we obtained system performance 76.48 and 71.33 BLEU score for English-Tigrigna vice versa respectively. But, if the training set and testing set are separated randomly, we obtained system performance 67.57 and 64.33 BLEU score for English-Tigrigna vice versa respectively.

Conclusion

The purpose of this study was to develop Bidirectional English-to-Tigrigna machine translation system using hybrid approach in which syntactic reordering rules are applied to reorder the structural difference between English and Tigrigna language pairs and finally the translation is made by using a statistical approach. This is an improvement on existing work on bidirectional English to Tigrigna translation using statistical (BiETSMT) approach carried out by [16] . Also, the system has a higher translation accuracy than Google translate and BiETSMT as well. However, there is still room for enhancements in this study. There is also need for a rule that can translate English words that have more than one meaning due to their part of speech. That is, a rule that will be able to recognize which translation will be appropriate for such words based on their part of speech in a sentence. The corpus size used for this study was 12,000 English-Tigrigna parallel sentences from four domains and 32,000 English-Tigrigna bilingual dictionaries are collected from scratch and incorporated for the sake of handling OOV. Collecting the corpus was difficult since there was not prearranged data for the bilingual English-Tigrigna corpus. Finally, hybrid approach in this system translation is still the most realizable method for translations shown from the result.

References

- [1] N. Kaur, V. Pushe, and R. Kaur, "Natural Language Processing Interface for Synonym," vol. 3, no. 7, pp. 638–642, 2014.
- [2] D. Liu, Y. Li, and M. A. Thomas, "A Roadmap for Natural Language Processing Research in Information Systems," *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, pp. 1112–1121, 2017.
- [3] P. Gupta, N. Joshi, and I. Mathur, "Quality Estimation of Machine Translation Outputs Through Stemming," *Int. J. Comput. Sci. Appl.*, vol. 4, no. 3, pp. 15–24, 2014.
- [4] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges," *IJCSI Int. J. Comput. Sci. Issues*, vol. 11, no. 5, pp. 159–165, 2014.
- [5] Y. K. Tedla, K. Yamamoto, and A. Marasinghe, "Tigrinya Part-of-Speech Tagging with Morphological Patterns and the New Nagaoka Tigrinya Corpus," vol. 146, no. 14, pp. 33–41, 2016.
- [6] "ATLAS - Tigrinya: The Tigrinya Language." [Online]. Available: <http://www.ucl.ac.uk/atlas/tigrinya/language.html>. [Accessed: 13-Sep-2018].
- [7] "ATLAS - Tigrinya: Welcome to the Tigrinya Taster Site!" [Online]. Available: <http://www.ucl.ac.uk/atlas/tigrinya/intro.html>. [Accessed: 13-Sep-2018].
- [8] "The Tigrinya Writing System." [Online]. Available: <http://www.ling.upenn.edu/courses/ling202/WritingSystem.html>. [Accessed: 14-Sep-2018].

- [9] I. Yamamoto, “The Development of Japanese-Uighur Machine Grammatical Comparison of,” *Language (Baltim).*, no. C, pp. 7–10, 2011.
- [10] C. Science and K. Sachdeva, “Hindi to English Machine Translation,” no. February, 2016.
- [11] O. . Ibrahim and Y. Mikami, “Stemming Tigrinya Words for Information Retrieval,” *Proc. COLInG 2012*, vol. 1, no. December, pp. 345–352, 2012.
- [12] C. Science, T. Performance, I. For, and L. Stemmer, “Towards Performance Improvement For Tigrigna,” 2016.
- [13] S. Stymne, “Pre-and postprocessing for statistical machine translation into germanic languages,” *Acl-2011*, no. June, pp. 12–17, 2011.
- [14] T. Hudík and A. Ruopp, “The Integration of Moses into Localization Industry,” *Proc. 15th Conf. Eur. Assoc. Mach. Transl.*, pp. 47–53, 2011.
- [15] C. Parra Escartín and M. Arcedillo, “A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings,” vol. 1, no. 2010, pp. 40–45, 2015.
- [16] M. Mulubrhan Hailegebreal, “College of Natural and Computational Sciences School of Information Science College of Natural and Computational Sciences,” 2017.