Statistical Machine Translator For English To Tigrigna Translation

Azath M., Tsegay Kiros

Abstract: Machine Translation is the automatic translation of text from a source language to the target language. The demand for translation has been increasing due to the exchange of information between various regions using different regional languages. English-Tigrigna Statistical Machine Translation, therefore, is required since a lot of documents are written in English. This research study used statistical machine translation approach due to it yields high accuracy and does not need linguistic rules which exploit human effort (knowledge). The language model, Translation model, and decoder are the three basic components in Statistical Machine Translation (SMT). Moses' decoder, Giza++, IRSTLM, and BLEU (Bilingual Evaluation Understudy) are tools that helped to conduct the experiments. 17,338 sentences of bilingual corpus for training, 1000 sentences for test set and 42,284 sentences for language model were used for experiment. The BLEU score produced from the experiment was 23.27% which would still not enough for applicable applications. As a result, the effect of word factored or segmentation in the translation quality is reduced by increasing the data size of the corpus.

Keywords: Machine Translation, Statistical Machine Translation, Bilingual corpus, and Monolingual Corpus.

1. INTRODUCTION

A language is a way of communication representing the ideas and expressions of the human mind. Hence the methodology of translation was adopted to communicate the messages from one language to another [1]. Developments in information communication and technology (ICT) have brought revolution in the process of machine translation [1]. Machine Translation is a subfield of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another [2]. The machine translation system, more specifically, is required to translate literary works from any foreign language into native languages. The foreign language text is usually fed into the machine translation system and the translation is done. Such systems can break language barriers by making available rich sources of literature to people across the world [2]. Statistical Machine Translation (SMT) is an approach of machine translation where a target sentence is generated on the basis of a large parallel corpus [3]. Statistical Machine Translation (SMT) is an approach to MT that is characterized by the use of machine learning methods [4]. In less than two decades, SMT has come to dominate academic MT research and has gained a share of the commercial MT market [4].

2. ABOUT TIGRIGNA LANGUAGE

Tigrinya is, a Semitic language of the Afro-Asiatic family that originated from the ancient Geez language, spoken in the East African countries of Eritrea and Ethiopia[5]. Ge'ez, is the ancient language, was introduced as an official written language during the first Axumite kingdom in Ethiopia when the Sabeans sought refuge in Axum[6]. It has still been being used by the Ethiopian Orthodox Tewahedo Church.

“ከንቲባከተማዓዲግራትኣይተተስፋልደትደስታብወገኖምህዝቢከተማዓዲግራትኩሎምመደባትልምዓትብዓቅሚመንግስቲክሰርሑከምዘይክእሉብምግንዛብብነፃዝተፈላለዩስራሕቲይሰርሕከምዘሎንፀጋማቱባዕሉናይምፍታሕባህሊእናዓብየይመፅእከምዘሎንኣብሪሆም፡” is a sample Tigrigna Text.

Unlike the Latin language, the Tigrigna script has more than 32 base letters with seven vowels each. Every first letter has six suffixes. Let’s have a look here:

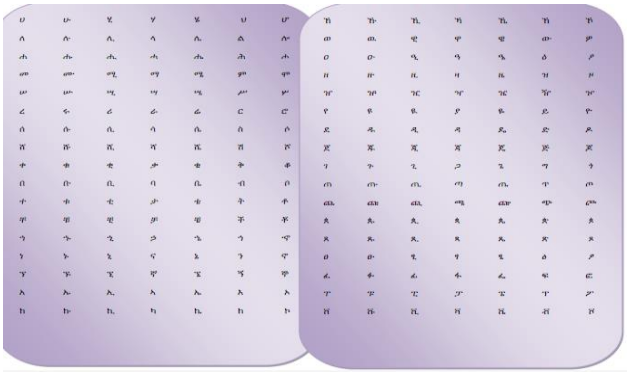

Figure 2: *Basic Geez Letters*

3. RELATED LITERATURE

Extensible researches have been being developed on the machine translation system. Here are some of the related works which have been reviewed from either international or local languages.Sinhal R. and Gupta K.[7] has designed a system of machine translation from English-Hindi using an approach of pure example-based machine translation. The researcher used 677 parallel English-Hindi corpuses for training and 150 non-parallel English sentences to test the precision. Researchers’ motivation toward this research was scarce in the availability of large-scale computational resources. The fundamental deficit of EBMT is it cannot satisfy user requirements as it is limited to some sample sentences. Ambaye T. and Yared M. [8] did research on an English to Amharic machine Translations System using an

- *Azath M Working as an Assistant Professor, Network & Security Chair, Faculty of Computing & Software Engineering, Arba Minch Institute of Technology, Arba Minch University, Ethiopia. His Research area with Network Security, Software Engineering and NLP. He Published many Articles in Springer and other reputed journals. E-mail: azathhussain@gmail.com*
- *Tsegay Kiros Working as a Lecturer, Programming Chair, Faculty of Computing & Software Engineering, Arba Minch Institute of Technology, Arba Minch University, Ethiopia. E-mail: tsegay_kiros@yahoo.com.*

approach of Statistical machine translation. The researchers have conducted five experiments collected from various sources using Moses toolkit. The study was evaluated using BLEU score in phrase-based translation model and Hierarchical translation model. Phrase-based BLEU score offered higher overall accuracy than hierarchical BLEU score. But, hierarchical translation model has higher performance in reordering than the phase translation model. Mulubrhan H. [9]has deployed on the research of Bidirectional Tigrigna-English Statistical Machine Translation. In this research Moses which is a free toolkit allowing automatic training for translation model using parallel corpus was used. During corpora preparation, data from different sources or domains were collected to make five sets of corpora. Experiments were conducted in phrase-based, morph-based and post-processed segmented comparatively. The researcher used a BLEU score in evaluating the performance of each set of experiments. In overall experiments, post-processed experiment outperformed baseline and morph-based experiments. And BLEU score gave higher accuracy for Tigrigna-English than its counterpart English-Tigrigna. Researcher, finally, concluded that corpus size and type during corpus preparation has an impact on translation quality.

4. DESIGN AND ARCHITECTURE

All important methods which have necessarily been included in accomplishing the study are presented as follows:

4.1 Data Collection

The quality of translation in statistical machine translation is dependent on the quality of data which have been fed to the Bilingual corpus and monolingual corpus. In general for training and testing, data is prepared from the bible found on the website: www. Geezexperience.com, FDRE constitution, Tigray Regional state constitution, high school textbooks, criminal law and news of sport, business and others from different mass Medias such as Tigray television, dimtsi weyane Tigray, Ethiopian broadcast Corporation, BBC(Britain Broadcast corporation) Tigrigna and VOA(voice of America).

4.2 The architecture of the Proposed System

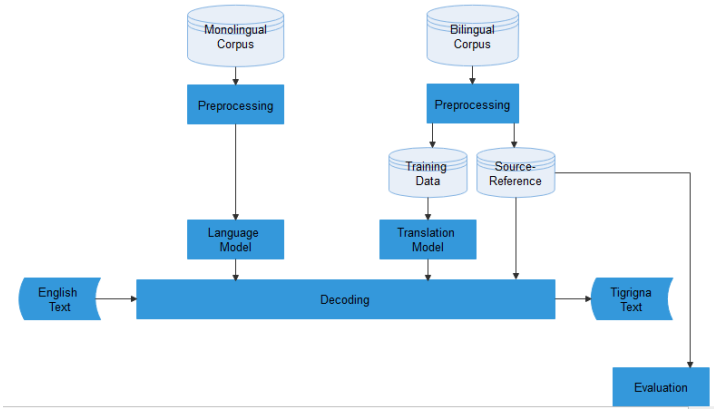

Figure3: *English-Tigrigna System Architecture*

4.3 Language Model

Language model estimates of how probable a sequence of words is to appear in the target language (Tigrigna in this case). In this research, the N-gram model is selected as it is the most widespread, simple and robust[10]. N-gram model can be defined as how likely words are to follow each other [9]. Therefore, the language model is calculated using an n-gram language which is computed from the monolingual corpus. The n-gram model can be unigram, bigram, trigram or higher-order n-grams. Let's have a look below on the target language (Tigrigna) sentences:

ፀፃይቤትትምህርቲከይዱ ።
ፀፃይቤትትምህርቲተፀዊዑ ።
ፀፃይቤትህንፀትተፀዊዑከይዱ ።
ፀፃይምሳሕበሊዑ ።
ትግርኛቋንቋትግራይእዩ ።

The bigram probability can be calculated by:

$$P(W2/W1) = \frac{\text{Count}(W1W2)}{\text{Count}(W1)}$$

$$P(\text{ቤት}/\text{ፀፃይ}) = \frac{\text{Count}(\text{ፀፃይቤት})}{\text{Count}(\text{ፀፃይ})} = \frac{3}{4} = 0.75$$

"ፀፃይ" and "ቤት" has been occurred together 3 times and 4 is the number of times the word "ፀፃይ" occurs.

4.4 Translation Model

The translation model is the most probable target language sentence given source language sentence P (E|T)[9][6]. An English-Tigrigna Bilingual corpus is prepared for the Translation model. In this model sentences of Source language, which is English, are aligned with sentences of Target language which is Tigrigna in this case in the training corpus. The probability of getting a target sentence t from a source sentence e can be calculated by:

$$P(t|e) = \frac{\text{count}(t,e)}{\text{count}(e)} \quad (4.1)$$

In the above equation, the sentence of English and Tigrigna are going to be partitioned into feasible words or phrases as it would be difficult to find a sentence of source in its counterpart target sentence.

$$P(e|t) = \sum_a p(a,e|t) \quad (4.2)$$

Variable a is the alignments between the individual chunks in the sentence pair.

The alignment probability of the chunks can be defined as:

$$p(a,t|e) = \prod_{j=1}^{m} t(tj,ei) \quad (4.3)$$

t(tj,ei) is the translation probability and can be computed as:

$$t(tj,ei) = \frac{\text{count}(tj,ei)}{\text{count}(ei)} \quad (4.4)$$

4.5 Decoding

The Moses decoder used the popular phrase-based decoder [11] and the Beam Search algorithm to find the best translation for the given input. For instance, translating a sentence e in the source language E (English in this case) to a sentence t in the target language T (Tigrigna in this case), the best translation is the probability that maximizes the product of the probability of English- Tigrigna

translation model p(e|t) and the language model of Tigrigna p (t), this is derived as follows below:

Bayes rule states that[12][4][13]:

$$P(t|e) = \text{argmax } P(e|t)*p(t) \quad (4.5)$$

Where P(e|t) and p(t) are the translation model and language model and e is source language and t is target language.

In mathematical terms[12][13]:

$$t_{best} = \text{argmax } p(t|e) \quad (4.6)$$

Substituting Equation 4.5 on Equation 4.6 produces Noisy channel model[13] which is described below.

$$t=\text{argmax } p(e|t)*p(t) \quad (4.7)$$

Where P(e|t) is the translation model, modeling the transformation probability from e to t and P(t) is the language model, assessing the overall well-formedness of the target sentence.

5. EXPERIMENTATION AND DISCUSSION

5.1 Bilingual Corpus

Bilingual Corpus, in this case, is one text document of source language English is put in parallel with its counterpart target language Tigrigna text document. The parallel corpus used total sentences of 17, 338.

Table5.1: *Bilingual Corpus*

Type of data	English		Tigrigna	
Units	Token	Sentence	Token	Sentence
Ethiopian Constitution	17,263	2080	8,076	2,080
Bible	97,182	13, 714	60,183	13,714
Tigray Constitution	14,234	1,650	7,019	1,650
Criminal code	5,147	734	4,237	734
Mass Media	3,254	432	2,156	432
Textbook	2,345	378	1,834	378
Total	139,425	17,338	83,505	17,338

5.2 Monolingual corpus

The data composition of the monolingual corpus is shown below.

Table 5.2: *Tigrigna corpus*

Data domain	Tigrigna	
	Tokens	Sentences
Ethiopian Constitution	8076	2080
Bible	112,543	29,754
Tigray Regional Constitution	7019	1650
Criminal code	13,231	2500
Mass Media	21,213	5342
Textbook	4367	958
Total	166,449	42,284

5.3 Training the system

Training of the system was made on the translation model, language model and decoding using GIZA++, IRSTLM, and Moses toolkits respectively. IRSTLM toolkit, therefore, was used for the target language Tigrigna. A Forward trigram language model, with Kneser-Ney as a smoothing tool was applied. The word alignment, phrase extraction, and scoring were used and lexicalized reordering tables and Moses configuration files were created with the training translation system. Mainly this step creates a “moses.ini” file, which is used for decoding and the phrase table is also created which basically contains the probabilities of a word following another.

5.4 An Experiment on Phrase-based Baseline system

The Phrase-based Baseline systems have been trained English-Tigrigna parallel training sets collected from different domain areas and tested using English source sentences which would give an output target Tigrigna. BLEU score was used to evaluate the output. The result obtained from the BLEU score is shown in the table below.

Table 5.4: *Phrase-based baseline BLEU Score*

Phrase-based Translation	Corpus Name	BLEU score
	Corpus	English-Tigrigna
		23.27%

A simple output of the experiment is shown here: English sentence “She is a student.” gave a Tigrigna sentence of “ንሳተምሃሪእዩ።”

5.5 Evaluation

METEOR, BLEU, NIST and TER are the most common evaluation metrics. BLEU is a precision-oriented metric in that it measures how much of the system output is correct, rather than measuring whether the references are fully reproduced in the system output[14]. BLEU reports a high correlation with a human judgment of quality and is one of the most popular metrics in the field[15]. In addition, it calculates scores for individual segments, generally sentences, and then averages these scores over the whole corpus for a final score. NIST, BLEU with some alteration, calculates how informative a particular n-gram is given more weight for correct rarer n-gram found on the translation and lower weight for more likely occurring n-gram. However, METEOR is designed to address some of the deficiencies inherent in the BLEU metric by including synonymy, a stemmer, part of speech, etc.[15]. In this research study, therefore, BLEU has been selected as it has a high correlation to human judgment and measures how much of the system output is correct. The output produced from a training set of 17,338 sentences which were collected from the Bible, FDRE constitution, criminal code, Mass Medias, etc. using BLEU score was 23.27%. 1000 sentences are used as a test set. The accuracy point of view and the time it takes to translate a particular sentence was the evaluation perspectives. The maximum time taken to translate English sentence to Tigrigna sentence is 2.394 seconds. The BLEU score produced is pictured below as a screenshot.

```
# ------------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        -------  -------  -------  -------  -------  -------  -------  -------  -------

BLEU: 0.3266    0.2876    0.2327    0.1812    0.1142    0.0734    0.0000    0.0000    0.0000    "moses"
MT evaluation scorer ended on 2018 Aug 28 at 16:27:32
================================================
```

***Figure** Error! No text of specified style in document.**1:** English-Tigrigna BLEU score*

6. CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

World Research Community on Machine Translation has been giving a prior to the Statistical machine translation approach from other common approaches such as example-based machine Translation and Rule-based machine Translation due to its requirement of limited Computational linguistic resources and giving a high translation quality. Statistical machine translation, in addition, is advised to languages that are morphologically rich like Tigrigna (resource-scarce). As a result, an experiment has been conducted in developing Statistical Machine Translator for English to Tigrigna translation.

Experiment		English-Tigrigna machine translations		
	Author	Mulubrhan H.	Yemane T.	Tsegay K(proposed)
Baseline	Average	15.22	15.6	23.27
	Maximum	20.55		
segmented	Average	17.14	20.9	

***Table 6.1:** Comparative Analysis with other's related study*

From table 6.1 we can observe that even if Mulubrhan H.[9]and Yemane T.[16] Used segmentation, the proposed system gave better results by increasing the corpus although it is not too much reasonable change or satisfactory. Finally, word segmentation or factored based have a great role in the quality of machine translation[9][17]. But, its role becomes insignificant as the size of both the bilingual and monolingual corpus gets increased. As the corpus increases the probability of the word being in different forms becomes high; and that was the role of factored or segmentation. In addition, since the content of corpus is dominated by data from Bible the accuracy of sentences selected from other domains in the test set would be affected. The overall result, which would have been produced, from 17,338 bilingual and 42,284 monolingual sentences collected from different domains using the BLEU score is 23.27%.

6.2 Recommendation

The following recommendations are forwarded for future work based on the finding of this study:

- Specific areas or domains like Tourism, Clinic, Meteorology, etc. would be advised.
- As the size of the corpus increases, the role of segmentation/factored is reduced. Therefore, better translation quality can be produced by increasing the training parallel and monolingual corpora.
- For local languages like Oromifa to Tigrigna and Amharic-Tigrigna is an open domain.

REFERENCES

[1] Juran krishna sarkhel sneha tripathi, "approaches to machine translation," Annals of Library and Information study, vol. 57, pp. 388-393, December 2010.

[2] Adetunmbi A.O, Oguntimilehin. A Abiola O.B, "A Review of the Various Approaches for Text to Text Machine Translations," International Journal of Computer Applications (0975 – 8887), vol. 120, no. 18, pp. 7-12, June 2015.

[3] Pranjal Das and Kalyanee K. Baruah, "Assamese to English Statistical Machine Translation Integrated with a Transliteration Module," International Journal of Computer Applications, vol. 100, pp. 20-24, August 2014.

[3] Pranjal Das and Kalyanee K. Baruah, "Assamese to English Statistical Machine Translation Integrated with a Transliteration Module," International Journal of Computer Applications, vol. 100, pp. 20-24, August 2014.

[4] Adam Lopez, "A Survey of Statistical Machine Translation," ACM Computing Surveys, vol. 40, pp. 1-49, August 2008.

[5] Yoshiki Mikami Omer Osman Ibrahim, "Stemming Tigrinya Words for Information Retrieval," in Proceedings of COLING 2012, Mumbai, 2012, pp. 345–352.

[6] Eleni Teshome, "Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus," Unpublished Masters Thesis, Departement of Computer Science, Addis Ababa University, March 2013.

[7]Ruchika A. Sinhal and Kapil O. Gupta, "A Pure EBMT Approach for English to Hindi Sentence Translation System ," I.J. Modern Education and Computer Science, vol. 7, pp. 1-8, July 2014.

[8]Ambaye Tadesse and Yared Mekuri, "aEnglish to Amharic Machine Translation Using SMT," The Prague Bulletin of Mathematical Linguistics, pp. 1-10, July 2012.

[9]Hailegebreal Mulubrhan, "Bidirectional Tigrigna – English Statistical Machine Translation," Unpublished master Thesis, School of Information Science, Addis Ababa University, june 2017.

[10] Amittai E. Axelrod, "Factored Language Models for statistocal Machine Translation," Unpublished Masters Thesis, University of Edinburgh, 2006.

[11]G.S. Vatsa, Nikita Joshi, and Sumit Goswami** Mukesh, "Statistical Machine Translation," DESIDOC Journal of Library & Information Technology, vol. 30, pp. 25-32, July 2010.

[12]Franz Josef Och and Daniel Marcu Philipp Koehn, "Statistical Phrase-Based Translation," proceedings of HLT-NAACL 2003 Main Papers, pp. 48-54, May-June 2003.

[13]Jianfeng Gao, "A Quirk Review of Translation Models," July 2011.

[14]Ngoc Phuoc An Vo and Octavian Popescu Simone Magnolin, "iLearning the Impact of Machine Translation Evaluation Metrics for Semantic Textual Similarity," In

Recent Advances In Natural Language Processing, pp. 398-403.
[15]R. B. Mishra Marwan Akee, "A Statistical Method for English to Arabic Machine Translation," International Journal of Computer Applications, vol. 86, no. 2, pp. 13-19, January 2014.
[16] Yemane Tedla and Kuzuhide Yamamoto, "The effect of shallow segmentation on English-Tigrinya statistical machine translation," in In Asian Language Processing (IALP), Tainan, 2016, pp. 1-18.
[17] Tariku Tsegaye, "English -Tigrigna Factored Statistical machine Translation," Unpublished masters Thesis, JUNE 2014.
[18] Fredrik Olsson, Atelach Alemu Argaw, Lars Asker Bj¨orn Gamb¨ack, "Methods for Amharic Part-of-Speech Tagging," in Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – aflat 2009, Athens, 2009, pp. 104–111.
[19] Jeff A. Bilmes and Katrin Kirchhoff, "Factored Language Models and Generalized Parallel Backoff," Association for Computational Linguistics, vol. 2, pp. 4-6, may-june 2003.
[20] M. Farr´us, J.B. Mari˜no, J.A.R. Fonollosa M.R. Costa-Juss`a, "STUDY AND COMPARISON OF RULE-BASED AND STATISTICAL CATALAN-SPANISH MACHINE TRANSLATION SYSTEMS," Computing and Informatics, vol. 31, pp. 245-270, 2012.