IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

Amharic-Awngi Machine Translation: An Experiment Using Statistical Approach

Habtamu Mekonnen

Dept. of Information science, Faculty of informatics, University of Gondar, Gondar, Ethiopia

Corresponding Author:hab.mekonnen@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i8.610 | Available online at: www.ijcseonline.org

Accepted: 09/Aug/2019, Published: 31/Aug/2019

Abstract- Nowadays, there are huge amount of texts written in Amharic language. Theses texts, documents and related Amharic literatures are usable for individuals, who can read, hear and understand Only Amharic languages. But In Ethiopia, there are many individuals who cannot hear and understand any text and literature written in Amharic language unless there is parallel translation in language they are good. These documents need to be translated to Awngi to provide valuable information for Awngi language speakers. To conduct the research, the corpus was collected from Amharic texts, Mass Media Agency and Bible. We used minimum of 1500 simple sentences, 1000 compound and 1000 complex sentences and maximum of 5000 sentences for each sentences type in order to train the system. We used 9:1 ratio for training and testing respectively. For language model we used minimum of 5700 and maximum of 14491 monolingual sentence of Awngi language. To do the system, we used Moses for Mere Mortal for translation process, MGIZA++ for alignment and IRSTLM for language model. Experimental results showed that better performance of 37% BLUE score was registered using complex sentences. In Amharic language, a word in sentences can have more than one meaning. While translating, the challenge of this study was not translating the meaning of the given sentences according to the context. But this study has not solved that challenge which needs further study to show all meanings of word depending on the context properly

Keywords—Awngi, Awngi translation, statistical, Amharic traslation, amharic-awngi translation

I. INTRODUCTION

Natural languages are those languages that are spoken by the people. Natural Language Processing (NLP) is the branch of computer science focused on developing systems that allow computers to communicate with people using everyday language [1]. The need for natural language processing is mainly due to a wide storage of information recorded or stored in natural language that could be accessible via computers. Information is constantly generated in the form of books, news, business and government reports, and scientific papers. A system requiring a great deal of information must be able to process natural language to retrieve much of the information available on computers [2]. The idea of computers being able to understand languages has been started since the first half of the twentieth century and was envisaged in a classic paper by Alan Turing (1950) as a hallmark of computational intelligence.

It is difficult for an individual to know and understand all the languages of the world. Machine translation (MT) is the task of automatically translating a text from one natural language into another [3]. In the modern world, there is an increased need for language translations due to the fact that language is an effective medium of communication. MT was one of the first envisioned applications of computers back in the 1950's. There are different reasons for conducting this research in machine translation. The most important reason is scientists, technologists, engineers, economists, agriculturalists, administrators, industrialists, businessmen, and many others have to read documents and have to communicate in languages they do not know. The second reason is enabling readers to understand some text, to grasp the general meaning of it.

Statistical machine translation (SMT) approach is a MT paradigm where translations are generated on the basis of statistical models. These statistical models parameters are derived from the analysis o bilingual text corpora. Statistical-based MT uses purely statistical based methods in aligning the words and generation of texts. SMT is based on the view that every sentence in a language has a possible translation in another language. It requires large sentence aligned parallel text for each language pair and this approach cannot be employed where these corpora are not available [4].

The direct approach is considered to be the most primitive or the original approach of all, carrying out replacement of the words in the source language with words in the target language. Another approach to machine translation is the rule-based approach to machine translation which involves the application of morphological, syntactic and semantic rules in the analysis of the source language text and the synthesis of the target-language text. RBMT parses the source text and produces an intermediate representation which may be a parse tree or some abstract representation. Human language, whether written or spoken, is a fundamental part of human communication. The only means by which human beings abstract reality is through language. Language is an efficient and effective medium of communication which explicitly represents the ideas and expressions of the human mind.

Amharic is a Semitic language, developed in the Horn of Africa during the 10th century [5]. It is the second-most spoken Semitic language in the world after Arabic. It is spoken as a first language by more than 27 million people of the country and official language of the Federal Democratic Republic of Ethiopia [6]. Awngi on the other hand is one of a Central Cushitic language family and spoken more than half a million people in a wide-area in Ethiopia.

There are many Amharic books (religious, historical, educational) that need to be translated. But there is no system that translates Amharic texts into Awngi. So the absence of studies to investigate these aspects in a professional context is the main problem. Because of this, people use human translation and they tend to be slower as compared to machines. Researchers believed that studying how to make these documents available in local languages (Awngi language) is vital in order to access valuable information from the collection. This thesis is organized in to five chapters. The first chapter discusses about introduction of the study: including statement of the problem, objective of the study, scope and limitation of the study, and methodology of the study. The second chapter discusses about literature review and language overview which focus on approach of machine translation with different tools used for corpus alignment and related works of the study, details of language relationship between two languages (Nature of Amharic and Awngi languages). The third chapter discusses about designing processes of the prototype including corpus preparation, types of corpus used for the study, and briefly discuss about the prototype of the system. Chapter four discusses about Experimental analysis of the study which includes different experiments and the results of the experiments with interpretation of findings. The conclusion and recommendation of the study has discussed in Chapter 5.

II. RELATED WORK

English-Afaan Oromo Machine Translation

This research was conducted by Sisay in April, 2009 at Addis Ababa University with the objectives of applying SMT systems on English – Afaan Oromo language pair by using parallel corpus and to identify the challenges that need a solution regarding the language pair. The research was experimented using statistical approach. In this research, experimentation of statistical machine translation of English to AfaanOromoo was conducted by varying number of N-grams and a score of 17.74% was found. At the end the researchers recommend that, that these tools and techniques used in this research should be applied for other languages in Ethiopia to help the speakers of the languages reap the benefits of getting documents available in English without renouncing their own language.

Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus

The thesis was conducted by Eleni at Addis Ababa University in March 2013.The objective of this study was to design and develop a bi-directional English-Amharic machine translation system using constrained corpus. Researchers used Statistical machine translation approach. The researchers have seen results from two perspectives, one from the accuracy point of view and the other from the time it takes to translate a particular sentence. Experiments were carried out based on the dataset and results were recorded. The experiments were taken separately, one for the simple sentences and the other for complex sentences. The result obtained for the simple sentence using BLEU Score had an average of 82.22% accuracy 20 for the English to Amharic, 90.59% for the Amharic to English and using the manual questionnaire preparation method, the accuracy from English to Amharic was 91% and from Amharic to English was 97%. For the complex sentences, the result acquired from the BLEU.

Score was approximately 73.38% for the English to Amharic, 84.12% for the Amharic to English and from the questionnaire method from English to Amharic was 87% and from Amharic to English was 89%. From this, we can see that the difference with the BLEU score and the questionnaire preparation method is not that visible so we can use both methods as reference. At the end of the work, the researchers have recommended that, further researches in machine translation on Amharic to other languages, in Ethiopia such as Tigrigna, AfaanOromo or so could be performed while preparing a large corpus.

English-Amharic Statistical Machine Translation

This was the research conducted by Mulu and Besacier [21]. The result recorded from the BLEU score methodology was 35.32%. At the end, researchers recommended that more experimentation and research is required to further improvement of the translation accuracy.

Bidirectional English-AfaanOromo Machine Translation
This was the research conducted by Jabesa [22] in 2013, with the objective of developing a bidirectional English-AfaanOromo machine translation system using hybrid approach. The result recorded from the BLEU score methodology was 32.39% for English to AfaanOromo translation and 41.50 % for AfaanOromo to English translation by using statistical approach. The researchers recommended that, the rules which are developed and used in the system are only used for syntax reordering. Therefore, additional results can be accomplished by further exploring the rules especially by developing morphological rules.

Optimal Alignment for Bi-directional AfaanOromo-English Machine Translation
This was the research conducted by Yitayew in june2017 by using statistical machine translation approach. The aim of the study was to explore the effect of word level, phrase level and sentence level alignment on bi directional AfaanOromo-English statistical machine translation. For the purpose of the system they used Moses for Mere Mortal for translation process, MGIZA++, Anymalign and hunalign tools for alignment and IRSTLM for language model. After preparing the corpus, different experiments have conducted. Experimental results showed that better performance of **47%** and **27%** BLUE score has registered using phrase level alignment with max phrase length 16 from AfaanOromo-English and from English-AfaanOromo translation respectively. This shows an improvement on the average 37 % accuracy registered in this study. According to researchers, the reason for this score was length of phrase level aligned corpus handle word correspondence. This represents that alignment has a great effect on the accuracy and quality of statistical machine translation in machine translation. Researchers in their study concluded that phrase level aligned corpus improves the performance of statistical machine translation, when the source and the target languages are English and AfaanOromo. Lastly, authors have identified as many-many alignment is a major challenge at phrase level that needs further investigation [37].

Awngi – Amharic Cross Language Retrieval
Esubalew has also conducted research by using Awngi language. His research title was "Developing Awngi-Amharic cross language information retrieval (CLIR)"by using dictionary based approach. From his study, the result obtained was 63%. Lastly, applying the same approach in phrases and compound words was put as future research direction by the authors of this research.

Steaming Algorithm for Awngi Text
Tsegaye has done thesis entitled as "Developing Steeming Algorithm for Awngi text" by using a long much approach. From his study the result scored was 91.4%. Finally, what they recommended for other researchers was to apply other algorithms and to see the result difference.

Sentiment Analysis Model for Opinionated Awngi text
This was the research conducted by Melese in June, 2017[24]. The main objective of the research was to develop feature level sentiment analysis models for opinionated Awngi language. Three classification techniques: Naive Bayesian, Maximum Entropy and Support Vector Machine algorithms has experimented for building and evaluating the model. To evaluate the performances of the systems; the precision, recall, f-score and accuracy were employed and model was constructed using those three algorithms. From the study, an experimental result that researchers achieved was average 75% accuracy for the model. As conclusion, researchers concluded that by examining factors that make the sentiment classification problems with Awngi language, there is no standardized corpus for sentiment mining. The results of the ML-based sentiment mining model for opinionated Awngi texts using the processes explained above are encouraging. However, further work can be done to improve the proposed model's results by standardized text corpus.

III. METHODOLOGY

In order to conduct the research, the researchers followed experimental research design hence different experiments are conducted to investigate performance of statistical machine translation by using different sentences types.

IV. RESULTS AND DISCUSSION

The main purpose of this study was to conduct experiment on Amharic –Awngi statistical machine translation to see sentences level translation for better performance of statistical machine translation.
The following table summarizes all experiments conducted and results obtained from experiments.

Table 1: Summary of Experiments

Types of sentences	Amount of sentences			Experiment	Performance
Simple	Training set	Test set	LM		10% (1-gram scoring) and 0.68% (3 gram scoring)
	1350	150	5700	Experiment I	
	4500	500	5700	Experiment II	11% (1-gram scoring) and 0.75% (3-gram scoring)
	4500	500	14491	Experiment III	11% (1-gram scoring) and 0.75% (3-gram scoring)
Compound	900	100	5700	Experiment I	3.6%
	4500	500	5700	Experiment II	4.56% and 0.25% for 1 gram and 3 gram ordering respectively
	4500	500	14491	Experiment III	4.56% and 0.22% for 1 gram and 3 gram ordering respectively
Complex	900	100	5700	Experiment I	11% (1-gram scoring) and 0.75% (3-gram scoring)
	4500	500	5700	Experiment II	37% (1-gram scoring) and 16.6% (3-gram scoring)
	4500	500	14491	Experiment III	37.28% (1-gram scoring) and 17.26(3-gram scoring)

As shown from the above result summary table, the sentences which performed best is complex sentences having training set of 4500 and test set 500 by using 14491 corpus for language model. Its record is maximum of 37.28% (1-gram scoring) and 17.26% (3-gram scoring) BLEU score. Increasing corpus for language model has shown performance improvement for complex sentences only. This is because the probability of complex sentences occurrence in language model is high. This increases the number of aligned sentences at phrase translation table and makes the translation performance high. For language model, we used mixed sentences of all type. For first three experiments (Experiment I- Experiment III), corpuses of source language used in language model training are similar with corpuses used for translation model training. But for fourth experiment (Experiment IV), corpuses of source language researchers used in language model training are different from corpuses used for translation model training. In addition to that identifying effect of language model and translation model for statistical machine translation is other basic strength of this study. As seen from experimental results translation model plays great role for performance improvement rather than language model. There for increasing corpus for translation model will give better result for machine translation than increasing corpus for language model. As we observed from the table, compound sentences performance was low in relative to other types of sentences. This is mainly due to the fact that the compound sentence is very large and complicated which makes it hard for the candidates to assess the translation process.

V. CONCLUSION AND FUTURE SCOPE

Nowadays in Ethiopia, there are huge amount of texts written in Amharic language. But there are many individual who cannot hear and understand text and literature written in Amharic language unless it is translated. These documents need to be translated to Awngi by the help of machine to provide that valuable information for Awngi language speakers. Therefore, machine translation plays an important role to handle language barriers between peoples and documents who want to access. The aim of this study was to translate Amharic sentences to Awngi sentences using statistical machine translation approach. Statistical MT tries to generate translations using statistical methods based on bilingual text corpora. In order to conduct the research, the researchers followed experimental research design. The researchers identified features and types of Amharic sentences, which are a good means to know statistical machine translation. Experiments were taken using simple sentences, compound sentences and complex sentences. The maximum result obtained for the simple sentence using BLEU Score was an average of 11.6% accuracy. For the complex sentences, the maximum result acquired was 37%. The result of 4.56 % was obtained for compound sentences. The result recorded was somehow high in complex sentences. During translation, output of the some input sentences was combination of both languages. This is one challenge that we faced during the study. The second challenge of the study is unable to get exact output for some translation inputs that are identical to given Amharic sentences. Generally this research concludes the finding of the study firstly: using complex sentences improves the performance of statistical machine translation, when the source and the target languages are Amharic and Awngi languages respectively. Secondly Translation model plays great role for performance improvement rather than language model.

- Experimentation and the addition of more bilingual data will raise the accuracy level of this system. Therefore, the researchers strongly recommend the addition of more bilingual data for further experimentation.
- For a given language, accuracy of translating from Amharic – Awngi and from Awngi to Amharic may be different. Thus, it is good to compare which direction of translation gives a better result using the given corpus.
- In this study we used statistical machine translation approach; researchers recommend applying other approach and to see its effect in performance.
- This study is sentences level translation excluding compound complex sentences type. So we recommend other researchers to conduct research by including this sentences type.
- Researchers recommend other researchers to come with the improvement of translation output.
- For language model training, we use the same corpus for all types of sentences. We recommend other researchers to conduct experiment by using simple sentences for language model to test simple sentences, to apply compound, and complex sentences for language model and test compound and complex sentences respectively.
- For this study, we used word level alignment, we recommend other researchers to conduct other experiment by using phrase level alignment and sentences alignment to get improved system performance.

ACKNOWLEDGMENT

First, Glory to God for all things. Then, we would like to thank our thesis advisor Dr. Adane Letta at University of Gondar for his excellent and enduring support. This thesis would not have been possible without his advice; continuous interest and shaping this work considerably and made the process of creating this work valuable learning experience. We would also like to thank Dr. Million Meshesha from the Addis Ababa University for his teaching about facts of research and research flow at the beginning. We would like to thank Abebe Belay (PhD candidate) from Taiwan University for his resource sharing. We want to give gratitude for Yitayew Solomon from Metu University hence

he helped more by sharing his experience on his previous work on machine translation by using Moses for Mere Mortals decoder .Our heartfelt gratitude goes to Asamirie Yemata, who helped by reading the work and gave constructive comment. Lastly and importantly, gratitude goes to family and all friends who helped during the study.

REFERENCES

[1]. Raymond J. Mooney,"CS 343: Artificial Intelligence Natural Language Processing", University of Texas , Texas, pp 1-4 ,2018.
[2]. Abhimanyu Chopra, Abhinav Prashar, Chandresh Sain. International journal of technology enhancements and emerging engineering research, vol 1, issue 4.pp.131 – 133, 2013
[3]. Karen Louise Smith. "The Translation of Advertising Texts: A Study of English Language Printed Advertisements and their Translations in Russian." PhD. thesis. 2002.
[4]. Esubalew Asmare Desta. "Developing Awngi-Amharic cross Language information retrieval (Clir):A Dictionary Based Query Translation Approach."M.Sc.thesis, University of Gondar, Ethiopia, 2015.
[5]. Sisay Adugna Chala. "English – Afaan Oromoo Machine Translation: An Experiment Using Statistical Approach." M.Sc.thesis, Addis Ababa University, Ethiopia, 2009.
[6]. Eleni Teshome." Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus."M.Sc.thesis, Addis Ababa University, Ethipia,2013.. Yitayew Solomon."Optimal Alignment for Bi directional Afaan Oromo-English Statistical MachineTranslation". Msc.Thesis, Addis Ababa University, Ethiopia,
[7]. June, 2017
[8]. Tsegeye Misikir,"Developing steeming algorithm for Awngi text: A longest much approach".Msc.Thesis, Addis Ababa University, Ethiopia, 2003
[9]. John Hutchins: Reflections on the history and present state of machine translation, University of East Anglia.
[10]. Llu´ıs M`arquez Villodre, "Empirical Machine Translation and its Evaluation", PhD.Thesis. Universitat Polit`ecnica de Catalunya, Barcelona, Maig de 2008.
[11]. John Hutchins,A new era in machine translation research: University of East Anglia, Norwich, England, pp.211-219], 1995.
[12]. Abiola O.B, Adetunmbi A.O, Oguntimilehin. A. A review of the various approaches for text to text machine translations, Afe Babalola University, Ado-Ekiti, Nigeria.
[13]. John Sturdy DeNero. "Phrase Alignment Mo dels for Statistical Machine Translation." PhD.thesis, University of California, Berkeley, 2010.
[14]. Jabesa." Bidirectional English-Afaan Oromo Machine Translation". Msc. Thesis. Addis Ababa University, Ethiopia, 2013.
[15]. Nakul Sharma, English to Hindi Statistical Machine Translation System, MSC thesis, Thapar University Patiala, June 2011.
[16]. Mulu, Besacier." English-Amharic Statistical Machine Translation". PhD. Thesis. Addisababa University, Ethiopia.
[17]. Ruchika Sinhal and Kapil Gupta, Language Processing for MT: Need, Problems and Approaches, International Journal of Engineering Research and General Science Volume 3, Issue 5, 2015.
[18]. Chris Callison-Burch , "Machine translation:benefits and advantages of statistical machine translation and NRC's Portage", pp. 1-14, 2005.
[19]. Melese Mihret." Sentiment Analysis Model for Opinionated Awngi Text".Msc. Thesis, University of Gondar. Ethiopia, 2017.
[20]. K. Dwivedi and P. P. Sukadeve, "Machine Translation System Indian Perspectives", Proceeding of Journal of Computer Science Vol. 6 No. 10. pp 1082-1087, May 2010.
[21]. M. D. Okpor. Machine Translation Approaches: Issues and Challenges. Journal, Vol. 11, Issue 5, No 2, September, 2014.
[22]. Khan Md. Anwarus Salam. , "Independent Study Report: Improving Example Based English to Bengali Machine Translation using WordNet", January 2009.
[23]. Mulu Gebreegziabher, Laurent Besacier. "English – Amharic Machine Translation: An Experiment Using Statistical Approach." PhD. Thesis, Addis Ababa University, Ethiopia.
[24]. John Sturdy DeNero. "Phrase Alignment Mo dels for Statistical Machine Translation", PhD.thesis, University of California, Berkeley,2010.

Authors Profile

Mr. Habtamu Mekonnen received his Bachelors of Science in information science from Jimma University and his Master of Science in computer science from University of Gondar. He is currently working at university of Gondar as lecturer in department of information science. His interest of research focuses on Machine learning technology, Internet of things, and big data science and analytics.

