Experimenting Statistical Machine Translation for Ethiopic Semitic Languages: The Case of Amharic-Tigrigna

Michael Melese Woldeyohannis(✉) and Million Meshesha

Addis Ababa University, Addis Ababa, Ethiopia
{michael.melese,million.meshesha}@aau.edu.et

Abstract. In this research an attempt have been made to experiment on Amharic-Tigrigna machine translation for promoting information sharing. Since there is no Amharic-Tigrigna parallel text corpus, we prepared a parallel text corpus for Amharic-Tigrigna machine translation system from religious domain specifically from bible. Consequently, the data preparation involves sentence alignment, sentence splitting, tokenization, normalization of Amharic-Tigrigna parallel corpora and then splitting the dataset into training, tuning and testing data. Then, Amharic-Tigrigna translation model have been constructed using training data and further tuned for better translation. Finally, given target language model, the Amharic-Tigrigna translation system generates a target output with reference to translation model using word and morpheme as a unit. The result we found from the experiment is promising to design Amharic-Tigrigna machine translation system between resource deficient languages. We are now working on post-editing to enhance the performance of the bi-lingual Amharic-Tigrigna translator.

Keywords: Under-resourced language · Amharic-Tigrigna
Semitic language · Machine translation

1 Introduction

Computers with the ability to understand human language contribute greatly to the development of more natural man-machine interfaces through better processing speeds and storage capacity [1]. Beside this, the advancement of ICT and the rise of the internet as a means of communication led to an ever increasing demand for Natural Language Processing (NLP). Among these applications, Machine translation (MT) is one, which refers to a process by which computer software is used to translate a text from one language to another [2]. The ideal aim of machine translation systems is to produce the best possible translation with minimal human intervention. Translation is not just word-for-word substitution rather it is a complex task that the meaning of source must be fully restored in the target holding grammar, syntax and semantics of the languages [3].

F. Mekuria et al. (Eds.): ICT4DA 2017, LNICST 244, pp. 140–149, 2018.
https://doi.org/10.1007/978-3-319-95153-9_13

Moreover, a translator must interpret and analyze all of the elements in the text and know how each word may influence another; this requires an extensive expertise as well as familiarity with source and target languages. For these, machine translation can follow rule-based, example-based, statistical-based or else machine learning approach. In this work statistical machine translation (SMT) have been applied.

According to world language [4], there are around 7,097 known living languages in the world and most of them are under resourced, Especially the African languages which contribute around 30% (2139) of the world language highly grieve from the lack of sufficient NLP resources. This is specially true for Ethiopic languages such as Amharic, Tigrigna and Afaan-Oromo among others.

Ethiopia being multilingual and multinational country, its constitution decrees that each nation, nationality and people has the right to speak, write and develop its own language. However, a lot of written documents, brochures, text books, magazines, advertisements and the web that are being produced in Amharic language than other. These would result in unbalanced production and distribution of material in different languages as an official and working language of Ethiopia.

Thus, to bridge the gap there is a need to develop a system that translate materials and documents into multiple languages, thereby ensuring effective information and knowledge sharing among the public at large as much as possible. In this paper an attempt is made to design a bi-directional statistical machine translation for semitic Amharic-Tigrigna language.

2 Ethiopic Language

Ethiopia has 89 languages which are officially registered in the country with up to 200 different spoken dialects [4–6]. Among these languages, this study consider semitic languages specifically Amharic and Tigrigna. This is because Amharic and Tigrigna are the second and the third widely spoken semitic languages in the world, next to Arabic. Unlike other Semitic languages, such as Arabic and Hebrew, both Amharic and Tigrigna uses a grapheme based writing system called fidel /fidälə/ which is written and read from left to right derived from Ge'ez /gə'əzə/ [7,8]. The majority of Amharic and Tigrigna speakers found in Ethiopia even though there are also speakers in a number of other countries, particularly Eritrea, Italy, Israel, Canada, the USA and Sweden.

The name Amharic (አማርኛ-amarəñña) comes from the district of Amhara (አማራ) region in northern Ethiopia, which is thought to be the historic center of the language being the official working language of the government of Ethiopia and some regional state such as Addis Ababa, Amhara and Southern Nations, Nationalities and People (SNNP). Whereas, Tigrinya (ትግርኛ/tigriññа/) is one of the language spoken by the Tigray people and serves as a working language of Tigray regional state of Ethiopia; it is also widely spoken in central Eritrea as an official languages.

Amharic language is spoken by more than 25 million with up to 22 million native speakers while Tigrigna has more than 7 million with up to 4 million native

speakers [4]. The following section discuss a review of Amharic and Tigrigna writing system in a view of preparing parallel text corpus for the development of Amharic-Tigrigna statistical machine translation system.

2.1 Amharic Writing System

Amharic symbols are categorized into four groups consisting of 276 distinct symbols [9,10]; core characters, labiovelar, labialized and labiodental. The detail category is presented in Table 1.

Table 1. Distribution of Amharic character set

Category	Character set	Order	Total
Core characters	33	7	231
Labiovelar	4	5	20
Labialized	18	1	18
Labiodental	1	7	7
Total			276

As shown in Table 1 Amharic has a total of 231 distinct core characters, 20 labiovelar symbols, 18 labialized consonants and 7 labiodental. The first category possess 33 primary characters each representing a consonant having 7 orders in form to indicate the vowel which follows the consonant to represent CV syllables. In the same way, labiodental category contains a character ቨ/v/ with 7 order borrowed from foreign languages and appears only in modern loan words like ቪዛ/viza/. Similarly, the labiovelar category contains 4 (ቀ/k'/, ኀ/h/, ከ/k/ and ገ/g/) characters with 5 orders that generates 20 distinct symbols. Furthermore, there are 18 labialized characters; for instance, ሏ/ l^{w}a/, ሟ/m^{w}a/, ሯ/r^{w}a/ and ሷ/s^{w}a/.

In Amharic writing, all the 276 distinct symbols are indispensable due to their distinct orthographic representation. In the machine translation task, we mainly deal with distinct words rather than with orthographic representation; Table 2 presents graphemes that have been normalized into common graphemes.

Table 2. List of normalized Amharic graphemes

Graphemes variants	Graphemes count	Equivalent graphemes	Normalized graphemes
ሀ, ሐ, ኀ, and ኸ	4	/h/	ሀ/h/
አ and ዐ	2	/ʔ/	አ/ʔ/
ሥ and ስ	2	/s/	ስ/s/
ፀ and ጸ	2	/ ts'/	ፀ/ts'/

Thus, among the given character set, different graphemes that generates same word have been normalized. Among the given 33 core character set, graphemes with multiple variants have to be normalized into their sixth order along with derivatives to generate equivalent graphemes as shown in Table 2. The selection of graphemes is made based on the usage of character in Amharic document. Thus, as a result of normalizing the seven orders of (ሀ, ሐ, ኀ, ኸ) to ሀ/h/, (ሐ, ዐ) to ሐ/ʔ/, (ሥ, ሰ) to ሰ/s/, (ፀ, ጸ) to ፀ/ts'/ there is a great reduction in one-to-many, many-to-one and many-to-many modelling for machine translation task.

2.2 Tigrigna Writing System

Tigrigna symbols are grouped into three different categories; consisting of 249 distinct symbol [6]; these are core characters, labiovelar and labiodental. The detail category of distinct symbols used in Tigrigna writing systems is presented in Table 3.

Table 3. Distribution of Tigrigna character set

Category	Character set	Order	Total
Core characters	31	7	217
Labiovelar	5	5	25
Labiodental	1	7	7
Total			249

Tigrigna has a total of 217 distinct core characters, 25 labiovelar symbols and 7 labiodental. The first category possess 31 primary characters each representing a consonant having 7 orders in form to indicate the vowel which follows the consonant to represent CV syllables. Whereas the second category contribute 5 labiovelar character with each representing five order. In the third category, like Amharic, Tigrigna possess one labiodental character with an order of 7.

Unlike Amharic writing system, all the 249 distinct Tigrigna symbols are indispensable due to their distinct orthographic representation and sound without overlapping. Unlike the Amharic script writing, Tigrigna does not require normalization as they do not provide the same meaning using different graphemes.

3 Data Preparation

One of the most fundamental resources for any statistical machine translation system is to have a parallel corpora. Collecting standardized and annotated corpora is one of the most challenging and expensive task [11]. This is specially true when working with under resourced languages. Unlike English, European languages (like French and Spanish) and Asian languages (like Japanese and Chinese) Amharic and Tigrigna can be considered as an under-resourced and technologically less supported languages that suffers from devising digital corpora.

For this research project, bible has been selected as a domain. The selection is made due to the existence of comparable Amharic-Tigrigna corpus and complex nature of expression it contains.

Data collected from web cannot be used directly for statistical machine translation. Thus, the corpus collected from web has been aligned to create parallel Amharic-Tigrigna sentence. Then, the corpus has been normalized, cleaned and segmented accordingly for both language. Finally, verified by linguist to have Amharic-Tigrigna parallel corpora. Beside this, all typing errors are corrected and further filtering done to overcome the problem that may arise as a result of one-to-many, many-to-one and many-to-many relationship due to orthographic variation that generates the same meaning.

This may be due to unnormalized, typing error and uncleaned text in the sentence. Hence, to normalize, clean the corpus and align at the sentence level after identifying the sentence boundary using Perl[1] and Python[2] as a programming languages.

The phrase based translation between the concepts in the source and target sentence greatly affect the statistical machine translation experiment. Figure 1 presents a sample one-to-one, one-to-many, many-to-one and many-to-many word level translation for Amharic-Tigrigna language.

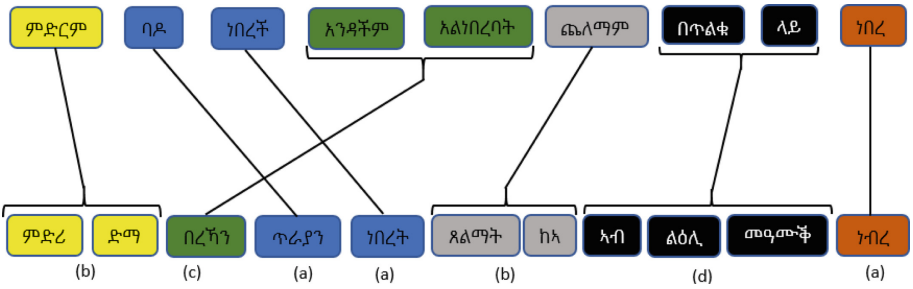

Fig. 1. Word correspondence between Amharic and Tigrigna (a) one-to-one, (b) one-to-many, (c) many-to-one and (d) many-to-many word translation mapping

Let us take a sample Amharic sentence "**ምድርም ባዶ ነበረች አንዳችም አልነበረባትም ጨለማም በጥልቁ ላይ ነበረ**" and its equivalent Tigrigna translation "**ምድሪ ድማ በረኻን ጥራያን ነበረት ጸልማት ከኣ ኣብ ልዕሊ መዓሙቕ ነብረ**". Before applying normalization, the Amharic word "**አንዳችም**" and "**አልነበረባትም**" has 2 variants **አ** and **ዐ**. This results in 4 (2^n where n is number character variant) possible combination in a given sentence of the of same meaning.

Similarly,
Tigrigna sentence translation of "**ምሸት ኮነ ብጊሓትውን ኮነ ሳልሰይቲ መዓልቲ**" has a

[1] Available at https://www.perl.org.
[2] Available at https://www.python.org.

total of 64 Amharic "**ማታም ሆነ ጥዋትም ሆነ ሦስተኛ ቀን**" sentence as a result one grapheme variation. Among this, 16 of them generated from **ሆ** variants whereas 4 from **ሦ** and **ስ** variants. Table 4 present sample unnormalized Tigrigna-Amharic sentence translation.

Table 4. Amharic-Tigrigna unnormalized translation

Tigrigna	Amharic
ምድሪ ድማ በረኻን ጥራያን ነበረት ጸልማት ከኣ ኣብ ልዕሊ መዓሙቕ ነብረ	ምድርም ባዶ ነበረች አንዳችም አልነበረባትም ጨለማም በጥልቁ ላይ ነበረ
	ምድርም ባዶ ነበረች አንዳችም ዐልነበረባትም ጨለማም በጥልቁ ላይ ነበረ
	ምድርም ባዶ ነበረች ዐንዳችም አልነበረባትም ጨለማም በጥልቁ ላይ ነበረ
	ምድርም ባዶ ነበረች ዐንዳችም ዐልነበረባትም ጨለማም በጥልቁ ላይ ነበረ
ይሁዳ ምስ ረኣያ ገጻ ሸፈና ነበረት እሞ ኣመንዝራ መሰለቶ	ይሁዳም ባያት ጊዜ ጋለሞታን መሰለችው ፊትዋን ተሸፍና ነበርና
	ይሑዳም ባያት ጊዜ ጋለሞታን መሰለችው ፊትዋን ተሸፍና ነበርና
	ይጉዳም ባያት ጊዜ ጋለሞታን መሰለችው ፊትዋን ተሸፍና ነበርና
	ይኹዳም ባያት ጊዜ ጋለሞታን መሰለችው ፊትዋን ተሸፍና ነበርና
	ይሁዳም ባያት ጊዜ ጋለሞታን መሠለችው ፊትዋን ተሸፍና ነበርና
	ይሑዳም ባያት ጊዜ ጋለሞታን መሠለችው ፊትዋን ተሸፍና ነበርና
	ይጉዳም ባያት ጊዜ ጋለሞታን መሠለችው ፊትዋን ተሸፍና ነበርና
	ይኹዳም ባያት ጊዜ ጋለሞታን መሠለችው ፊትዋን ተሸፍና ነበርና

Thus, more than 25,000 Amharic and Tigrigna sentences have been extracted from web in religious domain specifically bible. Then, the corpus were aligned to create a parallel corpora to fit the need of statistical machine translation with word and morpheme as a unit.

The Amharic corpus contains 25,470 sentences consisting of 355,993 tokens (64,259 types) with an average of 14 words per sentence. Similarly, Tigrigna corpus consist of 25,470 sentences consisting of 396,565 tokens (61,175 types) with an average of 16 words per sentence. Table 6 presents detail training, development, testing and language model data used for statistical machine translation

Table 5. Distribution of data per unit for Amharic-Tigrigna SMT

	Units	Amharic			Tigrigna		
		Sentence	Token	Type	Sentence	Token	Type
Training	Word	25,470	335,993	64,259	25,470	396,565	61,175
	Morpheme	25,470	541,425	23,809	25,470	561,057	30,138
Development	Word	500	7,362	3,374	500	8,917	3,015
	Morpheme	500	11,922	2,784	500	12,602	2,828
Testing	Word	1,000	13,845	6,042	1,000	16,300	5,439
	Morpheme	1,000	22,468	4,625	1,000	23,881	4,708
Language model	Word	36,989	679,716	112,511	62,335	1,089,435	109,988
	Morpheme	36,989	2,175,853	34,894	62,335	2,999,627	49,636

against the unit used for each language for words and morphemes parallel corpus (Table 5).

Beside the parallel corpus, we prepared a separate language model for both languages. The language model consists of 36,989 sentences (697,716 tokens of 112,511 types) for Amharic languages and 62,335 sentences (1,089,435 tokens of 109,988 types) for Tigrigna language at word level.

Similarly for the morpheme-based translation; the training, development, testing and language model data have been segmented into sub-word unit using corpus-based, language independent and unsupervised segmentation for both Amharic and Tigrigna language using morfessor 2.0[3] [12]. Figure 2 depicts the distribution of Amharic-Tigrigna training sentences used for word-based and morpheme-based translation.

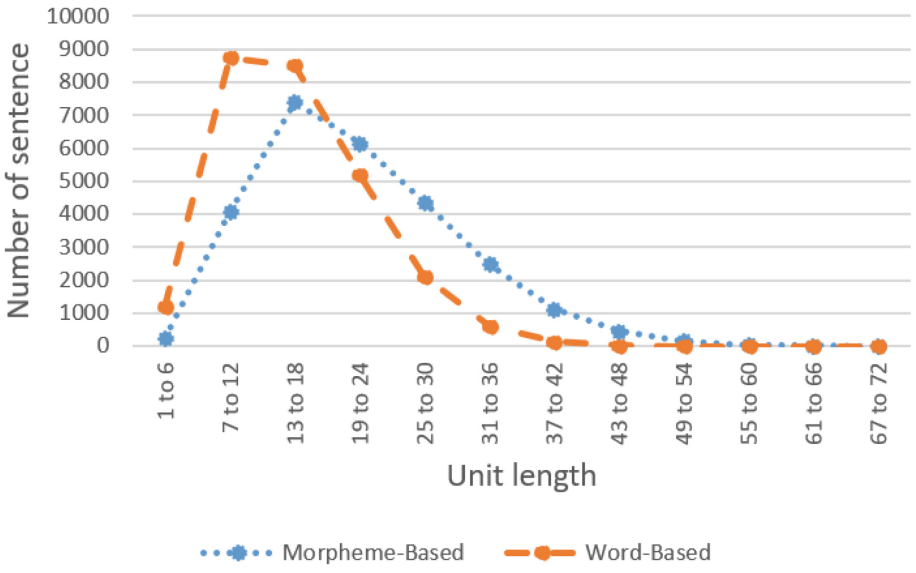

Fig. 2. Distribution of Amharic-Tigrigna sentence at word and morpheme levels

As we can see from Fig. 2, 67.8% of the parallel Amharic-Tigrigna machine translation data fall below 20 words per sentence. In addition to this, 6.7% of Amharic-Tigrigna sentence occurred more than 37 words per sentence.

4 Experiment

Both Amharic and Tigrigna are morphologically rich and complex languages; therefore, conducting the experiment through word and morpheme units

[3] The unit obtained with Morfessor segmentation is referred here as morpheme without any linguistic definition of morpheme.

are important. Towards putting the architecture inplace, a word and morpheme based translation have been conducted. For these we prepared a word-word, word-morpheme, morpheme-word and morpheme-based, word-based and morpheme-morpheme Amharic-Tigrigna and Tigrigna-Amharic parallel data. In addition to this, for each experiment, a word and morpheme based language model has been prepared for target language using tri-gram language model.

Once we prepared the parallel corpus for training, tuning and testing; Subsequently, the prepared parallel training data is aligned statistically by using multi-threaded giza (MGIZA). Then we trained the model using Moses and further tuned the model using the tuning data prepared for parallel corpus to create a translation model. A total of 8 models have been constructed taking word and morpheme as a unit of model. This includes, word-word, word-morpheme, morpheme-word and morpheme-morpheme for both Amharic-Tigrigna and Tigrigna-Amharic machine translation.

Once the translation model is ready for word and morpheme based translation, test data is prepared for evaluating the prototype. For this, 1000 sentences have been selected for word and morpheme based evaluation from the same domain. Then, the performance of each translation model have been tested at a sentence level using Bilinguale Evaluation Understudy (BLEU). Table 6 depicts the result obtained from machine translation with respect to each unit of translation.

Table 6. Distribution of data per unit for Amharic-Tigrigna SMT

Units			Target			
			Amharic		Tigrigna	
			Word	Morpheme	Word	Morpheme
Source	Amharic	Word			8.25	9.11
		Morpheme			9.09	13.49
	Tigrigna	Word	6.65	10.71		
		Morpheme	5.81	12.93		

Accordingly, the prototype have been evaluated using the same units (word-word and morpheme-morpheme) and different units (morpheme-word and word-morpheme). The word-word unit based translation correctly translated with BLEU score of 6.65 and 8.25 from Tigrigna-Amharic and Amharic-Tigrigna respectively. In addition, using morpheme as a unit for Amharic and Tigrigna, Amharic-Tigrigna resulted 13.49 while Tigrigna-Amharic scores 12.93 BLEU score.

On the contrary, a BLEU score of 5.81 for Tigrigna-Amharic and 9.11 for Amharic-Tigrigna have been achieved using word unit for Amharic and morpheme unit for Tigrigna. Moreover, using morph unit for Amharic and word unit for Tigrigna, 10.71 for Tigrigna-Amharic and 9.09 BLEU score for Amharic-Tigrigna have been achieved. Figure 3 presents summary of BLEU score registered for bilingual Amharic-Tigrigna statistical machine translation for each unit combination.

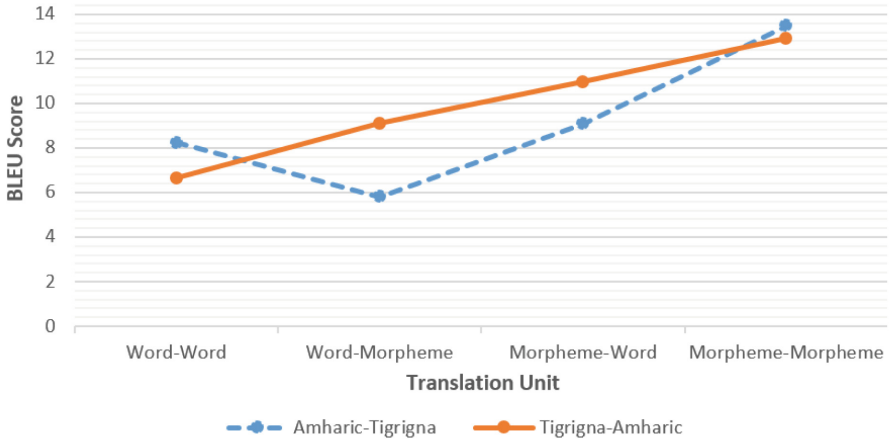

Fig. 3. Experimental result for bilingual Amharic-Tigrigna machine translation

Moreover, the performance of target side segmentation registered better result of translation than that of source side segmentation. The performance improvement in the target morpheme is as a result of minimizing morphological variation introduced in translation of the test set.

5 Concluding Remarks

In this research an attempt is made to design a bilingual machine translation for Amharic-Tigrigna. To conduct this study a total of 27,470 parallel Amharic and Tigrigna sentences have been selected from religious domain bible. The selected corpus have been preprocessed and analyzed morphologically using morfessor. Beside this, word and morpheme are used as translation unit with a sentence pair of maximum 80 words or morphemes per sentence selected after removing punctuation that should not be translated such as exclamation mark and colon.

Finally, the result obtained in this research is promising and serve as a proof that it is possible to have a SMT system for implementing a translation system for a pair of local languages. Experimental result shows that morpheme-based Amharic-Tigrigna translation outperforms word-based translation with a performance improvement by 4.24%. Beside this, SMT system may miss the real meaning of the source information since it depends on the size of corpus used for training. Had it been further analyzed beside morphological analysis, by combining with example based translation it would even give a better result than the current output. Hence, it is our next research direction to is to integrate example based with SMT system and further work on post-editing to enhance the performance of translator.

Acknowledgement. We would like to thank Ethiopia Ministry of Communication and Information Technology (MCIT) for funding to collect parallel text corpus and conduct an experiement for a bilingual Amharic-Tigrigna statistical machine translation research project.

References

1. Nakamura, S.: Overcoming the language barrier with speech translation technology. Sci. Technol. Trends Q. Rev. **31**, 35–48 (2009)
2. What is machine translation, SYSTRAN: we speak your industry's language. http://www.systran.co.uk/systran/corporate-profile/translation-technology/what-is-machine-translation
3. Martínez, L.G.: Human Translation Versus Machine Translation and Full Post-editing of Raw Machine Translation Output. Dublin City University, Dublin (2003)
4. Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the World, 20th edn. SIL, Dallas (2017)
5. Zekaria, S.: Summary and Statistical Report of the 2007 Population and Housing Census. Central Statistical Agency, Addis Ababa (2008)
6. Ager, S.: Omniglot, the online Encyclopedia of writing systems and languages
7. Hudson, G.: The world's major languages: Amharic. In: The World's Major Languages, 2nd edn, pp. 594–614. Routledge, Oxon/New York (2009)
8. Abyssinica dictionary: Amharic, the official language of Ethiopia (2015)
9. Teferra, S., Menzel, W., Tafila, B.: An Amharic speech corpus for large vocabulary continuous speech recognition. In: Proceedings of the XVth International Conference of Ethiopian Studies, Hamburg, Germany (2005)
10. Woldeyohannis, M.M., Besacier, L., Meshesha, M.: A corpus for Amharic-English speech translation: the case of tourism domain. In: Mekuria, F., Nigussie, E.E., Dargie, W., Edward, M., Tegegne, T., et al. (eds.) ICT4DA 2017. LNICST, vol. 244, pp. 129–139. Springer, Cham (2018)
11. Besacier, L., Le, V.-B., Boitet, C., Berment, V.: ASR and translation for under-resourced languages, Grenoble cedex 9, France
12. Creutz, M., Lagus, K.: Unsupervised discovery of morphemes. In: Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02, pp. 21–30, Philadelphia, Pennsylvania (2002)