CONSTRUCTION OF AMHARIC-ARABIC PARALLEL TEXT CORPUS FOR NEURAL MACHINE TRANSLATION

Ibrahim Gashaw and H L Shashirekha

Mangalore University, Department of Computer Science, Mangalagangotri, Mangalore-574199

ABSTRACT

Many automatic translation works have been addressed between major European language pairs, by taking advantage of large scale parallel corpora, but very few research works are conducted on the Amharic-Arabic language pair due to its parallel data scarcity. However, there is no benchmark parallel Amharic-Arabic text corpora available for Machine Translation task. Therefore, a small parallel Quranic text corpus is constructed by modifying the existing monolingual Arabic text and its equivalent translation of Amharic language text corpora available on Tanzile. Experiments are carried out on Two Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) based Neural Machine Translation (NMT) using Attention-based Encoder-Decoder architecture which is adapted from the open-source OpenNMT system. LSTM and GRU based NMT models and Google Translation system are compared and found that LSTM based OpenNMT outperforms GRU based OpenNMT and Google Translation system, with a BLEU score of 12%, 11%, and 6% respectively.

KEYWORDS

Amharic, Arabic, Parallel Text Corpus, Neural Machine Translation, OpenNMT

1. INTRODUCTION

Construction of a parallel corpus is very challenging and needs a high cost of human expertise. Machine Translation (MT), the task of translating texts from one natural language to another natural language automatically, is an important application of Computational Linguistics (CL) and Natural Language Processing (NLP). It can produce high-quality translation results based on a massive amount of aligned parallel text corpora in both the source and target languages [1]. MT systems need resources that can provide an interpretation/suggestion of the source text and a translation hypothesis. Parallel corpus consists of parallel text that can promptly locate all the occurrences of one language expression to another language expression and is one of the significant resources that could be utilized for MT tasks. [2].

The overall process of invention, innovation, and diffusion of technology related to language translation drive the increasing rate of the MT industry rapidly [3]. The number of Language Service Provider (LSP) companies offering varying degrees of translation, interpretation, localization, language, and social coaching solutions are rising in accordance with the MT industry [3]. Today many applications such as Google Translate and Microsoft Translator are available online for language translations.

There are different types of MT approaches, and many researchers have classified them in different ways [4] [5] [6]. Based on the rules and/or model used to translate from one language to another, MT approaches can be classified into Statistical Machine Translation (SMT), Rulebased

DOI : 10.5121/ijaia.2020.11107

Machine Translation (RBMT), Hybrid Machine Translation (HMT), and Neural Machine Translation (NMT). SMT uses a statistical model based on the analysis of large volumes of parallel text that determines the correspondence between a word from the source language and a word from the target language. RBMT uses grammatical rules that conduct a syntactic analysis of the source language and the target language to generate the translated sentence. However, RBMT requires extensive proof/reading and is heavily dependent on lexicons and domain experts. A hybrid approach is a combination of the statistical method and the rule-based approach, which includes a word-based model, a phrase-based model, a syntax-based model, and forest-based model. NMT is a type of MT that depends on neural network models to develop statistical models for translation.

Deep learning NMT approach is a recent approach of MT that produces high-quality translation results based on a massive amount of aligned parallel text corpora in both the source and target languages. Deep learning is part of a broader family of ML methods based on Artificial Neural Networks (ANN) [7]. It allows computational models that are composed of multiple processing layers to learn representations of data with various levels of abstraction. These methods have improved the state-of-the-art research in language translation [8]. NMT is one of the deep learning end-to-end learning approaches to MT that uses a large ANN to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. The advantage of this approach is that a single system can be trained directly on the source and target text no longer requiring the pipeline of specialized systems used in statistical MT. Many companies, such as Google, Facebook, and Microsoft, are already using NMT technology [9]. NMT has recently shown promising results on multiple language pairs. Nowadays, it is widely used to solve translation problems for many language pairs. However, much of the research on this area has focused on European languages despite these languages being very rich in resources.

Several MT systems have been developed, particularly from English to other natural languages, such as Arabic, Germany, Chinese, French, Hindi, Japanese, Spanish, and Urdu [1]. Though a limited amount of work has been done in different Ethiopian languages in the field of NLP, the MT system for Amharic-Arabic language pair is still in its infancy due to lack of parallel corpora. Therefore, it is essential to construct Amharic-Arabic parallel text corpora, which is very much required to develop the Amharic to Arabic NMT system.

Amharic language is the national language of Ethiopia spoken by 26.9% of Ethiopia's population as mother tongue and spoken by many people in Israel, Egypt, and Sweden. Arabic is a Semitic language spoken by 250 million people in 21 countries as the first language and serving as a second language in some Islamic countries. Ethiopia is one of the nations, which have more than 33.3% of the population who follow Islam, and they use the Arabic language to teach religion and for communication purposes. Both of these languages belong to the Semitic family of languages, where the words are formed by modifying the root itself internally and not merely by the concatenation of affixes to word roots [10].

NMT has many challenges, such as; domain mismatch, size of training data, rare words, long sentences, word alignment, and beam search [11] depending on the language pairs. Some of these challenges are addressed in this paper.

2. TRANSLATION CHALLENGES OF AMHARIC AND ARABIC LANGUAGES

Amharic and Arabic Languages are characterized by complex, productive morphology, with a basic word-formation mechanism, root-and-pattern. The root is a sequence of consonants, and the pattern is a sequence of Vowels (V) and Consonants (C) with open slots in it. It combines with the pattern through a process called interdigitating (intercalation): each letter of the root (radical)

fills a slot in the pattern. For example, the Amharic root s.b.r (sabr) denoting a notion of breaking, combines with the pattern CVCC (the slots and vowels are denoted by C and V respectively) [12].

In addition to the unique root-and-pattern morphology, they are characterized by a productive system of more standard affixation processes. These include prefixes, suffixes, infixes, and circumfixes, which are involved in both inflectional and derivational processes. Consider the Arabic word " وسوف يكتوبنـــــه " (wasawf yaktwubunahu) and its English translation "and they will write it". A possible analysis of these complex words defines the stem as "aktub" (write), with an inflectional circumfix, "y-uwna", denoting third person masculine plural, an inflectional suffix, "ha" (it), and two prefixes, "sa" (will) and "wa" (and). Morphological analysis of words in a text is the first stage of most natural language applications. Morphological processes define the shape of words. They are usually classified
into two types of processes [13];

1. A derivational process that deals with word-formation; such methods can create new words from existing ones, potentially changing the category of the original word. For example, from the Arabic root " كتـــب " (wrote), the following words are derived; " الكاتيـــــب" (the writer), " الكيتـــــآب " (the book), “ المكتبـــــة ” (the library), “ مكتبـــــه ” (library). The same is true in Amharic also. For example, from Amharic root “ፃ ፈ ” ” (he wrote), the following words are derived; "ፀ ሀ ፊ" (writer), "ጦፅ ሀ ፍ" (the book), "ቤ ተ ጦፅ ሃ ፍ" (the library).

2. Inflectional processes are usually highly productive, applying to most members of a particular word class. For example, Amharic nouns inflect for number, so most nouns occur in two forms, singular (which is considered in the citation form) and plural, regularly obtained by adding the suffix "ዎች" to the base form. This process makes the translation ambiguous. The considerable number of potential types of words and the difficulty of handling out-of-lexicon items (in particular, proper names) combined with prefix or suffix makes the computation very challenging. For example, in the word "aysäbramm", the prefix "ay" and the suffix "amm" (he doesn't break) are out-oflexicon items.

The main lexical challenge in building NLP systems for Amharic and Arabic languages is the lack of MRD which are vital resources. Absence of capitalization in Arabic and Amharic languages makes it hard to identify proper nouns, titles, acronyms, and abbreviations. Sentences in the Arabic language are usually long, and punctuation has no or little effect on the interpretation of the text.

Standard preprocessing techniques such as capitalization, annotation, and normalization cannot be performed on Amharic and Arabic languages due to issues of orthography. A single token in these languages can be a sequence of more than one lexical item, and hence be associated with a sequence of tags. For example, the Amharic word "አ ስ ፈ ” ረ ደ ች ብ ኝ " ("asferedachibegn"), when translated to English will be "a case she initiated against me was decided in her favor". The word is built from the causative prefix "as" (causes), a perfect stem "ferede" (judged), a subject maker clitics "achi" (she), a benefactive marker "b" (against) and the object pronoun "egn" (I).

Contextual analysis is essential in both languages to understand the exact meaning of some words. For example, in Amharic, the word "ገ ና " can have the meaning of "Christmas holiday" or "waiting for something until it happens." Diacritics (vowels) are most of the time, omitted from the Arabic text, which makes it hard to infer the word meaning and complex morphological rules should be used to tokenize and parse the text. The corpus of the Arabic language has a bias

towards religious terminology as a relatively high frequency of religious terms and phrases are found. Characters are sometimes stretched for justified text which hinders the exact match for the same word. Synonyms are very common in Arabic. For example, “year” has three synonyms عـسـام, حسـول, سســنه , and all are widely used in everyday communication.

Discretization is defined as a symbol over and underscored letters, which are used to indicate the proper pronunciations as well as for disambiguation purposes. Its absence in Arabic texts poses a real challenge for Arabic NLP, as well as for translation, leading to high ambiguity. Though the use of discretization is significant for readability and understanding, they don’t appear in most printed media in Arabic regions nor on Arabic Internet web sites. They are visible in the Quran, which is fully discretized to prevent misinterpretation [10].

3. RELATED WORKS

Various attempts have been made in the past to construct parallel text corpus for translation tasks from Amharic and Arabic languages to the other languages and translation models. Some of these works are as follows;

Tesfaye Bayu [14], describe the acquisition, preprocessing, segmentation, and alignment of Amharic-English parallel corpus that consists of 1,45,820 Amharic-English parallel sentences (segments) from various sources. This corpus is larger in size than previously compiled Amharic-English Bilingual Corpus (AEBC), which is hosted by European Language Resource Association with the size of 13,379 aligned segments (sentences) and the Low Resource Languages for Emergent Incidents (LORELEI) developed by Strassel and Tracey [15], that contains 60,884 segments.

Sakre et al., [16], presents a technique that aimed to construct an Arabic-English corpus automatically through web mining. The system crawled the host using GNU Wget in order to obtain English and Arabic web pages and then created candidate parallel pairs of documents by filtering them according to their similarity in the path, file name, creation date, and length. Finally, the technique measured the parallelism similarity between the candidate pairs according to the number of transnational tokens found between an English paragraph and its three Arabic neighbor paragraphs. However, in this work, they did not test or compare different models of statistical translation using the constructed parallel corpus.

Ahmad et al. [17] reported the construction of one million words English-Arabic Political Parallel Corpus (EAPPC) that consists of 351 English and Arabic original documents and their translations. The texts were meta-annotated, segmented, tokenized, English-Arabic aligned, stemmed, and POS-tagged. Arabic and English data were extracted from the official website of 'His Majesty King Abdullah II' and from 'His book' and then reprocessed from metadata annotation to alignment. They built a parallel concordancer, which consisted of two parts: the application through which the end-user interacts with the corpus and the database which stores the parallel corpus. Experiments carried out examined the translation strategies used in rendering a culture-specific term, and results demonstrated the ease with which knowledge about translation strategies can be gained from this parallel corpus.

Inoue et al. [18], describe the creation of Arabic-Japanese portion of a parallel corpus of translated news articles, which is collected at Tokyo University of Foreign Studies (TUFS). Part of the corpus is manually aligned at the sentence level for development and testing. The firstresults of Arabic-Japanese phrase-based MT trained on this corpus reported a BLEU score of 11.48.

Alotaibi [19], described an ongoing project at the College of Languages and Translation, King Saudi University, to compile a 10-million-word Arabic-English parallel corpus to be used as a resource for translation training and language teaching. The corpus has been manually verified at different stages, including translations, text segmentation, alignment, and file preparation, to enhance its quality. The corpus is available in XML format and through a user-friendly web interface, which includes a concordance that supports bilingual search queries and several filtering options.

All the above-related works use different lexical resources like Machine Readable Dictionaries (MRD) or parallel corpora, for probability assessment or translation. However, when there is a lack of such a lexical resource, an alternative approach should be available [20]. Nowadays, NMT Models are widely used to solve various translation problems. Learning Phrase Representations using Recurrent Neural Network (RNN) Encoder-Decoder for statistical MT [21] benefits more natural language-related applications as it can capture the linguistic regularities in multiple word level as well as phrase level. But it is limited to target phrases, instead of using a phrase table.

Dzmitry et al. [22], extended NMT Encoder-Decoder that encodes a source sentence into a fixed-length vector, which is used by a decoder to generate a translation. It automatically search for relevant parts of a source sentence to predict a target word without having to form these parts like a hard segment explicitly. Their method yielded good results on longer sentences, and the alignment mechanisms are jointly trained towards a better log-probability of producing correct translations that need high computational cost.

A. Almahairi et al. [23], proposed NMT for the task of Arabic translation in both directions (Arabic-English and English-Arabic) and compared a Vanilla Attention-based NMT system against a Vanilla Phrase-based system. Preprocessing Arabic texts can increase the performance of the system, especially normalization, but the model consumes much time for training.

4. CONSTRUCTION OF AMHARIC ARABIC PARALLEL TEXT CORPUS

As Amharic-Arabic parallel text corpora are not available for MT tasks, we have constructed a small parallel text corpus by modifying the existing monolingual Arabic and its equivalent translation of Amharic language text corpora available on Tanzile [24]. Quran text corpus consists of 114 chapters, 6236 verses, and 157935 words. The organization of the Quran text is categorized into verses (sequence of sentences and phrases). A sample verse in Arabic and its equivalent translation in Amharic and English is shown in Table 1. A total number of 13,501 Amharic-Arabic parallel sentences corpora have been constructed to train the Amharic to Arabic NMT system by splitting the verses manually into separate sentences of Amharic language as a source sentence and Arabic language as a target sentence as shown in Table 2. The total size of the corpus is 3.2MB, and it is split into training (80%) and test (20%).

In Arabic text corpus, as the first verse of each chapter of Quran start with " بِسْمِ اللَّسهِ الرَّحْمَٰنِ الرَّحِيمِ " (in the name of Allah the most gracious and the most merciful), it is split into a separate line. However, in case of Amharic corpus, the equivalent translation of this sentence is placed only in the first chapter. Therefore, it is added before the first line of the equivalent translated Amharic verses.

Table 1. Sample Verse of Quran in Arabic and its equivalent translation in Amharic and English

Chapter No: Verse No	Original Arabic Verses	Equivalent translation of Amharic Verses	Equivalent translat
2:282	يَا أَيُّهَا الَّذِينَ آمَنُوا إِذَا تَدَايَنتُم بِدَيْنٍ إِلَىٰ أَجَلٍ مُّسَمًّى فَاكْتُبُوهُ وَلْيَكْتُب بَّيْنَكُمْ كَاتِبٌ بِالْعَدْلِ وَلَا يَأْبَ كَاتِبٌ أَن يَكْتُبَ كَمَا عَلَّمَهُ اللَّهُ فَلْيَكْتُبْ وَلْيُمْلِلِ الَّذِي عَلَيْهِ الْحَقُّ وَلْيَتَّقِ اللَّهَ رَبَّهُ وَلَا يَبْخَسْ مِنْهُ شَيْئًا فَإِن كَانَ الَّذِي عَلَيْهِ الْحَقُّ سَفِيهًا أَوْ ضَعِيفًا أَوْ لَا يَسْتَطِيعُ أَن يُمِلَّ هُوَ فَلْيُمْلِلْ وَلِيُّهُ بِالْعَدْلِ وَاسْتَشْهِدُوا شَهِيدَيْنِ مِن رِّجَالِكُمْ فَإِن لَّمْ يَكُونَا رَجُلَيْنِ فَرَجُلٌ وَامْرَأَتَانِ مِمَّن تَرْضَوْنَ مِنَ الشُّهَدَاءِ أَن تَضِلَّ إِحْدَاهُمَا فَتُذَكِّرَ إِحْدَاهُمَا الْأُخْرَىٰ وَلَا يَأْبَ الشُّهَدَاءُ إِذَا مَا دُعُوا وَلَا تَسْأَمُوا أَن تَكْتُبُوهُ صَغِيرًا أَوْ كَبِيرًا إِلَىٰ أَجَلِهِ ذَٰلِكُمْ أَقْسَطُ عِندَ اللَّهِ وَأَقْوَمُ لِلشَّهَادَةِ وَأَدْنَىٰ أَلَّا تَرْتَابُوا إِلَّا أَن تَكُونَ تِجَارَةً حَاضِرَةً تُدِيرُونَهَا بَيْنَكُمْ فَلَيْسَ عَلَيْكُمْ جُنَاحٌ أَلَّا تَكْتُبُوهَا وَأَشْهِدُوا إِذَا تَبَايَعْتُمْ وَلَا يُضَارَّ كَاتِبٌ وَلَا شَهِيدٌ وَإِن تَفْعَلُوا فَإِنَّهُ فُسُوقٌ بِكُمْ وَاتَّقُوا اللَّهَ وَيُعَلِّمُكُمُ اللَّهُ وَاللَّهُ بِكُلِّ شَيْءٍ	እናንተ ያመናችሁ ሆይ! እስከ ተወሰነ ጊዜ ድረስ በዕዳ በተዋዋላችሁ ጊዜ ጻፉት። ጸሐፊም በመካከላችሁ በትክክል ይጻፍ። ጸሐፊም አላህ እንደ አስተማው መጻፍን እንቢ አይበል። ይጻፍም። ያም በርሱ ላይ ዕዳው ያለበት ሰው በኾሉ ያስጽፍ። አላህንም ጌታውን ይፍራ። ከእርሱም (ካለበት ዕዳ) ምንንም አያጉድል። ያም በርሱ ላይ ዕዳው ያለበት ቂል፣ ወይም ደከማ፣ ወይም በኾሉ ማስጻፍን የማይችል ቢኾን ዋቢው በትክክል ያስጽፍለት። ከወንዶቻችሁም ሁለትን ምስክሮች አስመስክሩ። ሁለትም ወንዶች ባይኾኑ ከምስክሮች ሲኾኑ ከምትወዱዋቸው የኾኑን አንድ ወንድና አንደኛዋ ስትረሳ አንደኛይቱ ሌላዋን ታስታውሰት ዘንድ ሁለት ሴቶች (ይመስክሩ)። ምስክሮችም በተጠሩ ጊዜ እንቢ አይበሉ። (ዕዳው) ትንሽ ወይም ትልቅ ቢኾንም እስከ ጊዜው ድረስ የምትጽፉት ከመኾን አትሰልቹ። እንዲህ ማድረጋችሁ አላህ ዘንድ በጣም ትክክል ለምስክርነትም አረጋጋጭ ለለመጠራጠራችሁም በጣም ቅርብ ነው። ግን በመካከላችሁ እጅ በጅ የምትቀበሏት ንግድ ብትኾን ባትጽፉዋት በናንተ ላይ ኃጢአት የለባችሁም። በተሸጣችሁም ጊዜ አስመስክሩ። ጸሐፊም ምስክርም (ባለ ጉዳዩ ጋር) አይጎዳ።	"O believers, when y a fixed term, draw writing, though bette scribe write it faithfu should refuse to wr him, and write what and have a fear of leave out a thing. If t of mind or infirm, o the guardian explain two of your men to two men are not avi two women you appr of them is confused her. When the wit they should not refus neglect to draw up a with the time fixed fo This is more equital and better as evidenc doubt. But if it i merchandise requiri face, there is no h

Table 2. Split Sentences of Quran Chapter 2 Verse number 282

Original Arabic Sentences	Equivalent translation of Amharic sentences	Equivalent transl sentences
يَا أَيُّهَا الَّذِينَ آمَنُوا	አናንተ ያመናችሁ ሆይ	O believers
إِذَا تَدَايَنْتُمْ بِدَيْنٍ إِلَى أَجَلٍ مُسَمًّى فَاكْتُبُوهُ	አበስ ተወሰነ ጊዜ ድረስ በሰዳ በተዋዋላችሁ	when you negotiate a
وَلْيَكْتُبْ بَيْنَكُمْ كَاتِبٌ بِالْعَدْلِ	ጸሐፊም በመካከላችሁ በትክክል ይጻፍ	though better it woul write it faithfully dow
وَلَا يَأْبَ كَاتِبٌ أَنْ يَكْتُبَ كَمَا عَلَّمَهُ اللَّهُ	ጸሐፊም አላህ አንደ አሰወቀው መጻፍን አንቢ አደባል	and no scribe should has taught him
فَلْيَكْتُبْ	ይጻፍም	and write
وَلْيُمْلِلِ الَّذِي عَلَيْهِ الْحَقُّ	ያም በርሱ ላይ ሰዳው ያለበት ሰው በቀሉ ያስጽፍ	what the borrower dic
وَلْيَتَّقِ اللَّهَ رَبَّهُ	አላህንም ጌታውን ይፍራ	and have fear of God,
وَلَا يَبْخَسْ مِنْهُ شَيْئًا	ከአርሱም ካለበት ሰዳ ምንንም አያጉድል	and not leave out a thi
فَإِنْ كَانَ الَّذِي عَلَيْهِ الْحَقُّ سَفِيهًا أَوْ ضَعِيفًا أَوْ لَا يَسْتَطِيعُ أَنْ يُمِلَّ هُوَ فَلْيُمْلِلْ وَلِيُّهُ بِالْعَدْلِ	ያም በርሱ ላይ ሰዳው ያለበት ቂል ወይም ደኸማ ወይም በቀሉ ማስጻፍን የማይችል ቢሾን ዋኒው በትክክል ያስጽፍለት	If the borrower is infirm, or unable to ex explain judiciously
وَاسْتَشْهِدُوا شَهِيدَيْنِ مِنْ رِجَالِكُمْ	ከወንዶቻችሁም ሁለትን ምስክሮች አስመስክሩ	and have two of witnesses
فَإِنْ لَمْ يَكُونَا رَجُلَيْنِ فَرَجُلٌ وَامْرَأَتَانِ مِمَّنْ تَرْضَوْنَ مِنَ الشُّهَدَاءِ أَنْ تَضِلَّ إِحْدَاهُمَا فَتُذَكِّرَ إِحْدَاهُمَا الْأُخْرَى	ሁለትም ወንዶች ባይሾኑ ከምስክሮች ሲሾኑ ከምትውዱዋቸው የሾኑን አንድ ወንድና አንደኛዋ ስትረሳ አንደኛዪቱ ሊለዋን ታስታውሳት ዘንድ ሁለት ሴቶች ይመስክሩ	but if two men are no and two women you a one of them is con prompt her
وَلَا يَأْبَ الشُّهَدَاءُ إِذَا مَا دُعُوا	ምስክሮችም በተጠሩ ጊዜ አንቢ አደባሉ	When the witnesses should not refuse (to c
وَلَا تَسْأَمُوا أَنْ تَكْتُبُوهُ صَغِيرًا أَوْ كَبِيرًا إِلَى أَجَلِهِ	ሰዳው ትንሽ ወይም ትልቅ ቢሾንም አስከ ጊዜው ድረስ የምትጽፉት ከመሾን አትሰልቹ	But do not neglect to or small, with the time the debt
ذَلِكُمْ أَقْسَطُ عِنْدَ اللَّهِ وَأَقْوَمُ لِلشَّهَادَةِ وَأَدْنَى أَلَّا تَرْتَابُوا	አንዚህ ማጽረጋችሁ አላህ ዘንድ በጣም ትክክል ለምስክርንትም አረጋጋጭ ላለመጠራጠራችሁም በጣም ቅርብ ነው	This is more equitabl and better as evidence doubt
إِلَّا أَنْ تَكُونَ تِجَارَةً حَاضِرَةً تُدِيرُونَهَا بَيْنَكُمْ فَلَيْسَ عَلَيْكُمْ جُنَاحٌ أَلَّا تَكْتُبُوهَا	ግን በመካከላችሁ እጅ በጅ የምትቀባበሏት ንግድ ብትሾን ባትጽፉዋት በናንተ ላይ ኃጢአት የለባችሁም	But if it is a deal ab requiring transaction f harm if no (contract is
وَأَشْهِدُوا إِذَا تَبَايَعْتُمْ	በተሸጣችሁም ጊዜ አስመስክሩ	Have witnesses to the

5. EXPERIMENTS AND RESULTS

In this work, we adopted openNMT Attention-based Encoder-Decoder architecture to construct Amharic Arabic NMT model, because attention mechanisms are being progressively used to enhance the performance of NMT by selectively focusing on sub-parts of the sentence during translation [25]. As described in [26], "NMT takes a conditional language modeling view of translation by modeling the probability of a target sentence w1:T given a source sentence x1:S as P(wt:T|x)=Π1 t P(wt|w1:t−1,x). This distribution is estimated using an Attention-basedEncoder-Decoder architecture". Two special kinds of Recurrent Neural Network (RNN) LSTM and GRU which are capable of learning long-term dependencies are used in this research work. RNN is a type of neural network for sequential data that can remember its inputs due to an internal memory which is more suited for machine learning problems. It can produce predictive results in sequential data that the information cycles through a loop when it makes a decision. It takes into consideration the current inputs and also previously received inputs, which is learned earlier [27].

LSTM was first introduced by S. Hochreiter and J. Schmidhuber [28], to avoid the long-term dependency problem. LSTM inherit the exact architecture from standard RNNs, with the exception of the hidden state. The memory in LSTMs (called cells) takes as input the previous state and the current input. Internally, these cells decide what to keep in and what to eliminate from the memory. Then, they combine the previous state, the current memory, and the input. LSTM calculates a hidden state ht as;

$$i_t = \sigma(x_t U_i + h_{t-1} W_i)$$

$$f_t = \sigma(x_t U_f + h_{t-1} W_f)$$

$$o_t = \sigma(x_t U_o + h_{t-1} W_o)$$

$$\widetilde{C}_t = \tanh(x_t U_{\widetilde{C}} + h_{t-1} W$$

$$C_t = \tanh(f_t * C_{t-1} + i_{t-1}$$

where t, i, f, o, W, U are called the time step, input gate, forget gate, output gate, the recurrent connection at the previous and current hidden layer, and the weight matrix connecting the inputs to the current hidden layers respectively. ˜Ct is a "candidate" hidden state that is computed based on the current input and the previous hidden state. C is the internal memory of the unit. GRU extends LSTM with a gating network generating signals that act to control how the present input and previous memory work to update the current activation, and thereby the current network state. Gates are themselves weighted and are selectively updated [29]. For GRU, the hidden state ht is computed as;

$$Z_t = \sigma(x_t U_z + h_{t-1} W_z)$$

$$r_t = \sigma(x_t U_r + h_{t-1} W_r)$$

$$\widetilde{h_t} = \tanh(x_t U_h + (r_t * h_{t-}$$

where, ~ht is activation function r is a reset gate, and z is an update gate. Both LSTM and GRU are designed to resolve the vanishing gradient problem which prevents standard RNNs from learning long-term dependencies through gating mechanism.

A basic form of NMT consists of two components; an encoder which computes a representation of source sentence S and a decoder which generates one target word at a time and hence decomposes the conditional probability [25]. The Attention-based Encoder-Decoder architecture used for Amharic-Arabic NMT is shown in Figure 1.

In general, the proposed model works as follows:

1. 1.Reads the input words one by one to obtain a vector representation from every encoder time step using LSTM/GRU based encoder

2. 2.Provide the encoder representation to the decoder

3. 3.Extract the output words one by one using another LSTM/GRU based decoder that is conditioned on the selected inputs from the encoder hidden state of each time step.

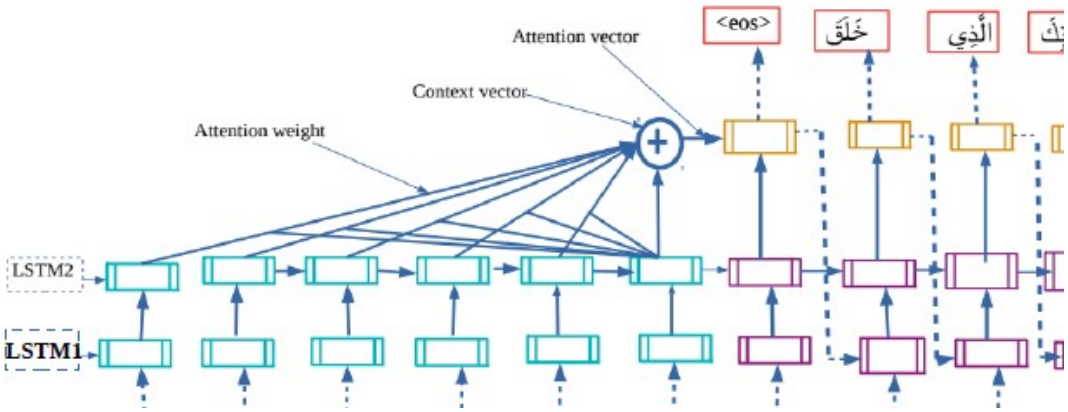

Figure 1: Attention based Encoder-Decoder architecture for Amharic-Arabic NMT

With this setting, the model is able to selectively focus on useful parts of the input sequence and hence, learn the alignment (matching segments of original text with their corresponding segments of the translation).

As shown in Figure 1, Using LSTM Attention based Encoder-Decoder, each word from the source sentence is associated with a vector $w \epsilon R^d$ and will be transformed into $[w_0, w_1, w_2, w_3, w_4] \epsilon R R^{dx5}$ by the encoder, and then an LSTM over this sequence of vectors is computed. This will be the encoder representation (attention weights) $e = [e_0, e_1, e_2, e_3, e_4]$. The attention weights and word vectors of each time step is fed to another LSTM cell to compute the context vector which is computed as:

$$h_t = LSTM(h_{t-1}, [w_{i-1}$$

$$s_t = g(h_t)$$

$$p_t = softmax(s_t)$$

where, g is a transformative function that outputs a vocabulary size vector. A soft-max is then applied to s_t to maximize it to a vector of probability $pt \epsilon R^V$. Each entry of p_t will measure how Figure 1: Attention based Encoder-Decoder architecture for Amharic-Arabic NMT LSTM1 likely is each word in the vocabulary and the highest probability pt is taken as $i_t = \text{argmax}(p_t)$ and and corresponding vector $w_{it-1} = w_{it}$

The attention or context vector Ct is computed at each decoding step first with the function $f(h_{t-1}, e_t') \rightarrow \alpha_t' \epsilon R$ and then a score for each hidden state e_t' of the encoder is computed. The sequence of α_t' is normalized using a soft-max and Ct′ is computed as the weighted average of e_t' as;

$$\alpha_{t'} = f(h_{t-1}, e_{t'}) \in R \text{ for}$$

$$\bar{\alpha} = softmax(\alpha)$$

$$C = \sum_n$$

The same procedure is also applied for GRU based NMT.

Preprocessing of both Amharic and Arabic scripts will have a great impact on the performance of the NMT system. Sentences are split and aligned manually and then all punctuation marks are removed from texts. After extensive experiments, the maximum source and target sequence length are set to 44, maximum batch size for training and validation is set to 80 and 40 respectively and learning rate to 0.001 with Adam optimization for both LSTM and GRU RNN type. The remaining parameters are used as default. The system saves the model for each of 10,000 training samples and then computes accuracy and perplexity of each model 10 times. Perplexity is a measure of how easily a probability distribution (the model) predicts the next sequence. A low perplexity indicates that the translation model is good at predicting/translating the test set. $PP(W)=P(w_1w_2.....wn)-1n$

Table 3 and Table 4 shows LSTM-base and GRU-based NMT evaluation, where, ”Val ppl”, ”Val. Accuracy”, ”Av. pred score” and ”Pred ppl” represents Validation Perplexity, Validation Accuracy, Average Prediction Score and Prediction Perplexity respectively. The result indicate that LSTM-based NMT outperforms GRU-based NMT. Since this is the first experiment done on Amharic and Arabic parallel text corpus we consider it as a good performance with small size corpus.

Table 3. LSTM-based NMT Evaluation

Epochs	BLEU-Score	Val. PPL	Val. Accuracy	Av. Pred. Sc
1	0.11	12725	33.21	-0.49
2	0.11	41181.5	33.50	-0.40
3	0.11	100996	33.64	-0.35
4	**0.12**	**100417**	**34.34**	**-0.34**
5	0.12	99881.1	34.32	-0.34
6	0.12	99876.1	34.33	-0.34
7	0.12	99876	34.33	-0.34
8	0.12	99876	34.33	-0.34

Table 4. GRU-based NMT Evaluation

Epochs	BLEU-Score	Val. PPL	Val. Accuracy	Av. Pred. Sc
1	0.108	13647	32.65	-0.51
2	0.101	65598.4	32.68	-0.39
3	0.098	172950	32.38	-0.35
4	0.105	173231	33.10	-0.34
5	**0.105**	**175635**	**33.12**	**-0.34**
6	0.105	175701	33.11	-0.34
7	0.105	175702	33.11	-0.34
8	0.105	175702	33.11	-0.34

The models are evaluated using Bilingual Evaluation Understudy (BLEU), which is a score for comparing a candidate translation of the text with reference translations. The primary programming task for a BLEU implementer is to compare n-grams of the candidate with ngrams

of the reference translation and count the number of matches which are position independent. More the matches, better the candidate translation is. BLEU is inexpensive to calculate and it is quick to use. It is expressed as the following equation [30];

$$BLEU = P_B exp\left(\sum_{n}^{N} w_n l\right.$$

where Pn is an n-gram precision that uses n-grams up to length N and positive weights wϵRn that sum to one.

We also compared the two recurrent units LSTM and GRU based OpenNMT translation algorithm with Google Translation System which is a free multilingual translation system developed by Google to translate multilingual text [9] and the results are shown in Figure 2. LSTM based OpenNMT outperforms over GRU based OpenNMT and Google Translation system, which is BLEU score of 12%, 11%, and 6% respectively.

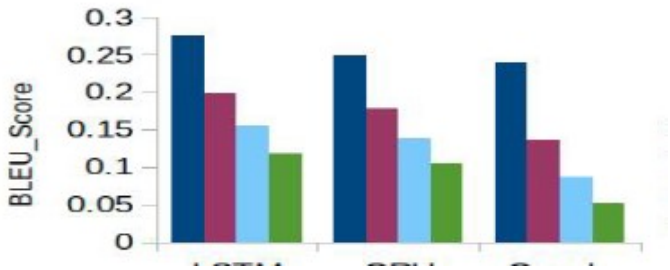

Figure 2: Best BLEU-Scores of LSTM and GRU based OpenNMT translation and Google Translation System

6. CONCLUSION

Many researchers have investigated to solve translation problems for many language pairs and NMT has recently shown promising results on multiple language pairs. However, much of the research on this area has focused on European languages despite these languages being very rich Figure 2: Best BLEU-Scores of LSTM and GRU based OpenNMT translation and Google Translation System in resources. Since Amharic and Arabic languages lack parallel corpus for the purpose of developing NMT, small size Amharic-Arabic parallel text corpora have been constructed to train the Amharic to Arabic NMT system by splitting the verses manually into separate sentences of Amharic language as a source sentence and Arabic language as a target sentence. Using the constructed corpus LSTM-based and GRU-based NMT models are developed and evaluated using BLEU. The results are also compared with Google Translation system. Since this is the first experiment done on Amharic and Arabic parallel text corpus, we consider it as a good performance for small size corpus. Extensive experiments with a large amount of training data could be implemented for better performance.

REFERENCES

[1] S. Islam, A. Paul, B. S. Purkayastha, and I. Hussain, "CONSTRUCTION OF ENGLISH-BODO PARALLEL TEXT CORPUS FOR STATISTICAL MACHINE TRANSLATION," Int. J. Nat. Lang. Comput. Vol, vol. 7, 2018.

[2] L. Rura, W. Vandeweghe, and M. Montero Perez, "Designing a parallel corpus as a multifunctional translator's aid," in XVIII FIT World Congress= XVIIIe Congrès mondial de la FIT, 2008.

[3] B. Maylath, "Current trends in translation," Commun. Lang. Work, vol. 2, no. 2, pp. 41–50, 2013.

[4] J. Oladosu, A. Esan, I. Adeyanju, B. Adegoke, O. Olaniyan, and B. Omodunbi, "Approaches to Machine Translation: A Review," FUOYE J. Eng. Technol., vol. 1, pp. 120–126, 2016.

[5] T. O. Daybelge, "Improving the precision of example-based machine translation by learning from user feedback," bilkent university, 2007.

[6] S. Ghosh, S. Thamke, and others, "Translation of Telugu-Marathi and vice-versa using rule based machine Translation," arXiv Prepr. arXiv1406.3969, 2014.

[7] MemoQ, "5 Translation Technology Trends to Watch Out for in 2019," 2019. [Online]. Available: https://slator.com/press-releases/5-translation-technology-trends-to-watch-out-in- 2019/.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, p. 436, 2015.

[9] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv Prepr. arXiv1609.08144, 2016.

[10] H. L. Shashirekha and I. Gashaw, "DICTIONARY BASED AMHARIC-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL," Comput. Sci. \& Inf. Technol., vol. 6, pp. 49–60, 2016.

[11] P. Koehn and R. Knowles, "Six challenges for neural machine translation," arXiv Prepr. arXiv1706.03872, 2017.

[12] M. Wordofa, "Semantic Indexing and Document Clustering for Amharic Information Retrieval," AAU, 2013.

[13] M. El-Haj, "Multi-document arabic text summarisation," University of Essex, 2012.

[14] Gezmu Andargachew Mekonnen Andreas Nurnberger and T. B. Bati, "A Parallel Corpus for Amharic-English Machine Translation," 2018.

[15] S. M. Strassel and J. Tracey, "LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages.," in LREC, 2016

[16] M. M. Sakre, M. M. Kouta, and A. M. N. Allam, "automated construction of Arabic-English parallel corpus," J. Adv. Comput. Sci., vol. 3, 2009.

[17] A. A.-S. Ahmad, B. Hammo, and S. Yagi, "ENGLISH-ARABIC POLITICAL PARALLEL CORPUS: CONSTRUCTION, ANALYSIS AND A CASE STUDY IN TRANSLATION STRATEGIES," Jordanian J. Comput. Inf. Technol., vol. 3, no. 3, 2017.

[18] G. Inoue, N. Habash, Y. Matsumoto, and H. Aoyama, "A Parallel Corpus of Arabic-Japanese News Articles," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[19] H. M. Alotaibi, "Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching," Arab World English J. Vol., vol. 8, 2017.

[20] P. Resnik and N. A. Smith, "The web as a parallel corpus," Comput. Linguist., vol. 29, no. 3, pp. 349–380, 2003.

[21] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv Prepr. arXiv1406.1078, vol. 3, 2014.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv Prepr. arXiv1409.0473, vol. 7, 2014.

[23] A. Almahairi, K. Cho, N. Habash, and A. Courville, "First result on Arabic neural machine translation," arXiv Prepr. arXiv1606.02680, vol. 1, 2016.

[24] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS.," in Lrec, 2012, vol. 2012, pp. 2214–2218.

[25] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv Prepr. arXiv1508.04025, 2015.

[26] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," arXiv Prepr. arXiv1701.02810, 2017.

[27] N. Donges, "Recurrent Neural Networks and LSTM," 2018. [Online]. Available: https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv Prepr. arXiv1412.3555, 2014.

[30] K. Wolk and K. Marasek, "Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts," Procedia Comput. Sci., vol. 64, pp. 2–9, 2015.

AUTHORS

Ibrahim Gashaw Kassa is a Ph.D. candidate at Mangalore University Karnataka State, India, since 2016. He graduated in 2006 in Information System from Addis Ababa University, Ethiopia. In 2014, he obtained his master's degree in Information Technology from the University of Gondar, Ethiopia., and he serves as a lecturer at the University of Gondar from 2009 to May 2016. His research interest is in Cross-Language Information Retrieval, Machine translation Artificial Intelligence Natural Language Processing.

Dr. H L Shashirekha is a Professor in the Department of Computer Science, Mangalore University, Mangalore, Karnataka State, India. She completed her M.Sc. in Computer Science in 1992 and Ph.D. in 2010 from University of Mysore. She is a member of Board of Studies and Board of Examiners (PG) in Computer Science, Mangalore University. She has several papers in International Conferences and published several papers in International Journals and Conference Proceedings. Her area of research includes Text Mining and Natural Language Processing.