Amharic to English Machine Translation Using Recurrent Neural Network

Dawit Assefa
HiLCoE, Computer Science Programme, Ethiopia
dawitassefa2013@gmail.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

Amharic is a morphologically rich language spoken by an estimated 30 million people in Ethiopia and in different parts of the world. Translating Amharic content into English has a very immense contribution on facilitating the communication among people. However, translating such content manually is very tedious, costly, and time-consuming. Machine Translation (MT) is an efficient approach to translate text without any human involvement. Neural Machine Translation (NMT) is a new and promising approach that has a remarkable success and become a competitive alternative to statistical MT approach. In this paper, we used NMT using sequence to sequence (seq2seq) Recurrent Neural Network (RNN) for translating Amharic content into English. Broadly experiments are divided into two: Baseline (one NMT model that doesn't use Attention mechanism) and Attention-Based (two NMT models with attention mechanism). A total of 95,846 sentence pairs were collected, pre-processed and used for training the models. Evaluation is performed in two ways: Automatic (BLEU score) and Manual (Human Evaluation). The experimental results confirm that attention based models (Attention_Bahdanu with 32.8% BLEU score and Attention_GNMT with 34.9% BLEU score) outperform the Baseline no_attentional model. Based on fluency and adequacy parameters, human evaluation results also confirm that attention-based models outperform the Baseline no attentional model.

Keywords: Neural Machine Translation; Recurrent Neural Network; Long Short Term Memory; Attention; BLEU Score

1. Introduction

Machine Translation (MT) is a sub-field of Natural Language Processing (NLP) which attempts to automate all or part of the process of translating from one human language to another. MT is perhaps the most challenging Artificial Intelligence (AI) task. Classically rule-based systems were used which were replaced with Statistical Machine Translation (SMT). More recently, deep neural network MT models achieve state-of-the-art results.

Amharic is morphologically rich language. Developing automatic MT system of Amharic texts to English has a very immense contribution on facilitating the communication between people. It could benefit the business industry to promote products to English speakers, to disseminate indigenous knowledge captured using Amharic to help tourists understand the history of Ethiopia and to translate relevant business and legal documents. Therefore, it is a good contribution if there is a way in which Amharic texts are automatically translated into English.

Though there are a few research works done on Amharic-English MT using statistical approaches, a new promising approach to MT, i.e., Neural Machine Translation (NMT), has a remarkable success and becomes a competitive alternative to the statistical approach. Therefore, having all the beauty of NMT features and no prior research conducted using neural based approach, this paper explores the development of deep learning neural network model for translating Amharic text into English. In this paper, design science research method is used. Specifically, neural based MT using seq2seq Recurrent Neural Network (RNN) is applied. RNN is selected because it is suitable for processing variable-length sequential data like sentences.

2. Related Work

Recently NMT is becoming a promising approach to MT in which a large network is trained to maximize the translation performance. Research on machine translation involving Amharic has a short history. In 2012, the work in [1] tried to implement a rule based machine translation based on Extensible Dependency Grammar that relies on constraint satisfaction for parsing and generation. The paper introduced a new framework for Rule Based Machine Translation (RBMT) and applied to the development of an English-Amharic MT system.

In 2012, Mulu Gebreegziabher and L. Besacier [2] conducted a preliminary experiment on English-Amharic Machine Translation using statistical approach. They used SRILM toolkit and GIZA++ for language modelling and word alignment, respectively. They used Moses decoder for decoding and achieved the base-line phrase-based BLEU score of 35.32%. By applying morpheme segmentation to the tokens of the Amharic training set, test set and the output of target text set of the baseline system, they achieved 36.58%. It was a very good start in terms of SMT for English to Amharic.

Three years later, in 2015, same authors with similar corpus [3] conducted a phoneme-based English-Amharic SMT focusing on improving the translation quality by applying phonemic transcription on the target side (Amharic). They achieved 37.53% BLEU score and they concluded that phoneme-based translation outperforms the baseline system.

In 2013, Eleni Teshome [4] conducted a study on bidirectional English-Amharic Machine translation using constrained corpus. This research was also implemented using SMT approach. Two experiments were conducted separately, one for simple sentences and the other for complex sentences. The work has seen results from accuracy point of view. The BLEU score achieved for simple sentences is 82.22% for English-Amharic and 90.59% for Amharic-English. Using the manual questionnaire method, accuracy of English to Amharic translation is 91% and accuracy of Amharic to English translation is 97%. Whereas, for complex sentences the BLEU score obtained was 73.38% for English-Amharic and 84.12% for Amharic-English and using manual questionnaire method, accuracy of English to Amharic translation is 87% and Amharic to English translation is 89%. It is concluded that MT translation from Amharic to English has a better accuracy than English to Amharic and given a large corpus, a proper result will be produced with a suitable time limit.

In 2016, Almahairi *et al*. [5] conducted end-to-end neural machine translation for bidirectional Arabic to English MT. The aim was to present the first result on the Arabic-English and English-Arabic bidirectional MT and they extensively compared vanilla attention-based neural MT against a vanilla phrase-based SMT. They classified the data into three; namely training set, in-Domain Evaluation set and out-of-domain evaluation set. For phrase-based SMT, they used Moses MT system as decoder, GIZA++ for word alignment and KenLM to build language model. For NMT, they used attention-based neural machine translation with bidirectional RNN-GRU (RNN with Gated Recurrent Unit) encoder and unidirectional RNN-GRU decoder. For English-Arabic translation, using in-domain evaluation set, they achieved 35.98% and 33.62% BLEU score and using out-of-domain evaluation set, they achieved 19.27% and 24.46% BLEU score for phrase-based and neural MT, respectively. Whereas, for Arabic-English translation, they used in-domain evaluation set only and they achieved 51.19% and 49.7% BLEU scores for phrase-based and neural MT, respectively. They concluded that performing proper pre-processing improves translation quality and phrase-based SMT and NMT system perform comparably to each other.

3. The Proposed Solution

3.1 Approach

In this paper deep NMT approach is used. This approach is based on neural networks and conditional probability of translated sentences from the source language to the target. The source language is Amharic and the target is English. The architecture of this study is based on the sequence-to-sequence (seq2seq) encoder-decoder architecture. To do machine translation with seq2seq enc-dec architecture, we connected RNN encoder with RNN

decoder language model. For this research, we used RNN with LSTM neural network. The RNN encoder processes the sentence in the source language word by word and generates a representation of the input sentence. The RNN decoder (or language model) takes the output from the encoder as input and generates the translation of the input sentence word by word [6, 7]. The proposed architecture of Amharic to English translation NMT model with attention is shown in Figure 1. The proposed architecture has five components: Pre-processing, Encoder, Decoder, Attention, and Evaluation. In pre-processing component, the collected parallel corpus (Amharic and English) is pre-processed through different methods such as normalizing, tokenizing, cleaning and subword segmentation using byte pair encoding (BPE) [8] for the representation of an open vocabulary through fixed-size vocabulary of variable-length character sequence. Then we shuffled and split the corpus into three sets: training, development and testing set. From the training set, vocabulary files are learning automatically and generated for both languages.

The Encoder component contains source embedding layer and 4 LSTM-RNN layers. The first RNN layer is Bidirectional (BiRNN) and the others are unidirectional (UniRNN). The task of the encoder is to provide a representation of the input sentence, which is a sequence of words. Given the categorical nature of words, the encoder first looks up the source embedding to retrieve the corresponding word representation. Then it processes these words with a LSTM-RNN network. This results in hidden states that encode each word with its preceding words. Also to get the right context (succeeding words), BiRNN encodes right-to-left so that we could have full sentence context representation.

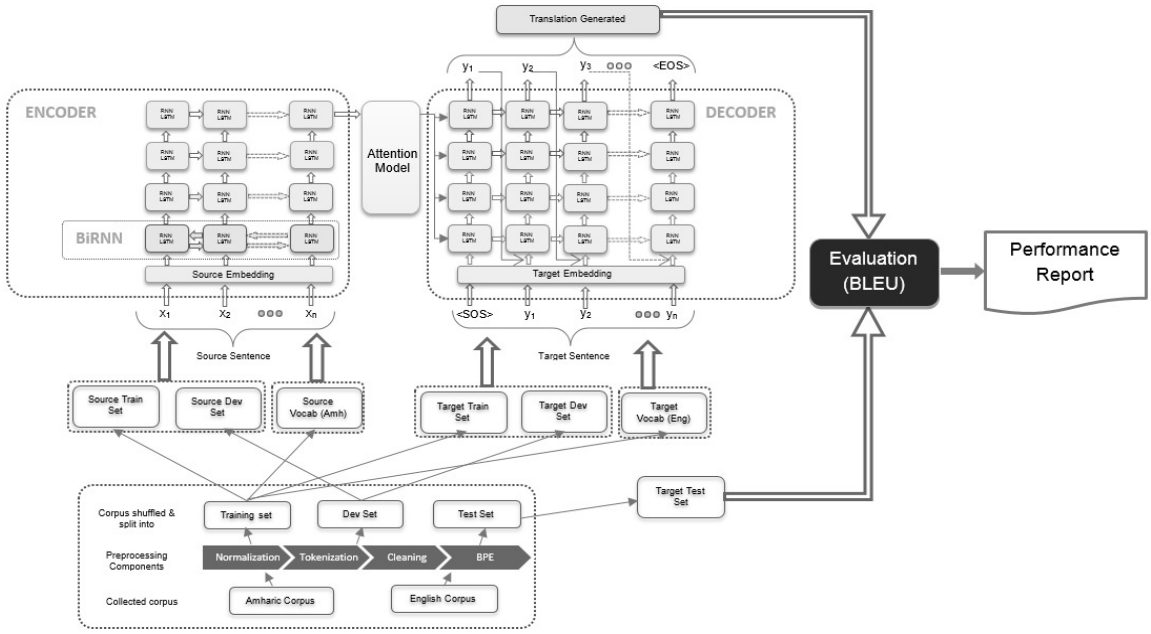

Figure 1: The Proposed Architecture of Amharic to English NMT Model with Attention

Like the Encoder component, the Decoder component uses LSTM-RNNs. It contains four unidirectional (UniRNN) layers. It takes the representation of the input context, the embedding of the previously selected word and the previous hidden state and it predicts the output word. This prediction takes the form of a probability distribution over the entire output vocabulary.

The Attention component ties the encoder and decoder components. For the attentional model experiments, we used attention mechanism [9]. This mechanism is informed by all input word representations and the previous hidden state of the decoder and it produces an input context state. The key idea of attention mechanism is to establish direct shortcut connections between the target and the

source by paying attention to relevant source contents during translation.

The last component is Evaluation of the system results compared with the reference data (target test set). Evaluation is performed in both ways: Automatic (BLEU) score and Manual Evaluation.

3.2 Data Collection and Preparation

A total of 95,846 sentence pairs were collected and pre-processed. We collected the parallel data from different sources as shown in Table 1.

Table 1: Collected Corpus by Domain

Domain	*# Sentence Pairs*	*%*
Jehovah Witness's publications	33,804	35.27
Bible (King James Version)	28,828	30.08
News and History	20,487	21.38
Proclamation	12,054	12.58
Constitution	598	0.62
UNHR	74	0.08

From the total, 65.35% of the data are from the Bible and related topics, 21.38% from News and History, 12.58% from Proclamation, 0.62% from the Constitution, and 0.08% from UN declaration of Human Rights.

The collected documents were combined, pre-processed and shuffled, then divided into *training set*, *development set*, and *testing set*. Since neural based machine translation needs large amount of parallel data for training the model, we used 98:1:1 ratio to give 98% of the total data for training and the rest 1% for development and 1% for testing.

3.3 Experimental Settings

The experiment in this study is broadly divided into two: *Baseline* and *Attention-Based*. The Baseline experiment consists of an encoder and a decoder with both RNN. The attention-based experiment consists of an encoder RNN, a decoder RNN, and attention mechanisms [9]. In Baseline enc-dec architecture, the encoder reads the whole sentence, memorizes it and stores it in the final activation layer, then the decoder network generates the target translation. The Attention-Based enc-dec architecture uses attention-mechanism. This mechanism decides how much attention should be paid to a particular word while translating the sentence.

Totally three experiments were conducted. *Experiment I* (*Baseline_noAttention*) explored and evaluated the performance of the baseline seq2seq enc-dec model without deploying attention in translating Amharic sentences to English. *Experiments II* (*Attention_Bahdanau*) and *III* (*Attention_GNMT*) were conducted to explore the attention mechanism with seq2seq enc-dec architecture [7] and Google Neural Machine Translation (GNMT) architecture [10], respectively and we evaluated their performance compared with *Experiment I* model. All the experiments were conducted on Linux machine with Linux Mint 18.3 Cinnamon 64 bit OS. Python combined with Tensorflow[1] (a large-scale deep learning system), and Anaconda (the Python data science platform) were used. Each experiment was conducted on *tf-seq2seq*[2] NMT framework which is developed by Google (Google's Artificial Intelligence Research and Development).

3.4 Hyper-Parameter Selection

The hyper parameter for each model is selected following the benchmark attention experiment [11]. The reason is, when starting a new application in deep learning, it is almost impossible to correctly guess the right value for all hyper parameters on the first attempt. So in practice, the process is highly iterative by defining initial hyper parameter values, code and experiment and get the result, and go through this cycle many times to find a good choice of hyper parameters for the application. This process is very expensive and time consuming. Therefore, we selected the hyper parameter settings for each model following the benchmark in [11]. But, during training of each model, we were constrained by computational resources. So we had to reduce some of the hyper-parameter values with respect to the hardware capacity used for implementation.

Therefore, for three of the experiments, the models are trained on the training data consisting of 95,846 Amharic to English sentence pairs. The

[1] https://www.tensorflow.org/

[2] https://github.com/google/seq2seq

vocabulary size is limited to 30,000 most frequent words on both languages. All the out-of-vocabulary (OOV) words are mapped to a special token unknown (UNK) and all get the same embedding. We used a LSTM network layer for both encoder and decoder RNN. We used four LSTM encoder layers and four LSTM decoder layers. For the first layer of encoder, we used bi-directional RNN-LSTM (BiRNN). The word embedding dimension and hidden unit numbers are set to 512. Limiting maximum sentence length is a key parameter to reduce memory usage. Therefore, the sentence length for the training dataset is set up to 50 tokens. As used in [11], dropout strategy [12] is applied to the output layer with the dropout rate of 0.2 to avoid over-fitting. The model is optimized using Stochastic Gradient Descent (SGD) with batch size of 128. Once the models are trained, beam search is employed to find the best translation with beam width 10. Seq2seq architecture without attention mechanism was used for *Experiment I* (*Baseline_noAttention*). Whereas, for *Experiment II* (*Attention_Bahdanau*), it adopts seq2seq architecture and unlike *Experiment I*, it uses Bahdanau's attention mechanism [9]. *Experiment III* (*Attention_GNMT*) also employs Bahdanau's attention mechanism. But the architecture is slightly different from the standard seq2seq architecture. It uses Google Neural Machine Translation (GNMT) architecture [10]. The GNMT model originally has one bidirectional RNN layer followed by seven unidirectional RNN layers. The decoder has eight unidirectional RNN layers [10]. However, to build such deep RNN layers is computationally very expensive. So, for the sake of this experiment, we need to limit the layers to 1 bi-directional followed by 3 uni-directional encoder layers and 4 decoder layers are used to train. In addition to using GNMT architecture, we applied subword segmentation on training set and fed to the GNMT architecture. Learning subword units from the given corpus improves the performance of MT by alleviating the open vocabulary issues [8].

4. Result and Discussion

The performance of each experiment was evaluated in two ways: Automatic evaluation (BLEU Score) and Human Evaluation. The performance of each experiment is presented in Tables 2 and 3.

Table 2: BLEU Score

Experiment Type	*BLEU*
Baseline_noAttention	24.5
Attention_Bahdanau	32.8
Attention_GNMT	34.9

From Table 2, we can observe that, Attention_Bahdanau NMT system exceeded the Baseline_noAttention system by 8.3 BLEU points. Whereas, Attention_GNMT system exceeded by 2.1 BLEU points over Attention_Bahdanau system and 10.4 BLEU points over the Baseline_noAttention experiment. These results show that incorporating attention yields higher quality translation. From Experiment III result, we can conclude that in addition to attention mechanism, using sub word segmentation techniques improve the machine translation performance by effectively handling the open-vocabulary issues.

On the other hand, in order to gain further insight, we employed human evaluation. The translation of 20 sentences randomly generated by each experiment was given to three people. They were selected based on purposeful sampling method. The most widely used methodology when manually evaluating MT is based on the parameter of fluency and adequacy [13]. Fluency refers to the grammatical accuracy of the translated text. Adequacy is defined as the degree to which the reference sentence is conveyed in the translation. Fluency and adequacy for a given translation have different levels on which they can be evaluated. The levels are given in Tables 3 and 4.

Table 3: Level of Fluency

Rank	*Parameter*	*Definition*
1	Nonsense	The translated sentence is not clear
2	Acceptable	The translated sentence is broken, but is understandable with efforts
3	Fair	The translated sentence is easy to understand but has lack correct grammar
4	Perfect	The translated sentence is good in grammar

Table 4: Level of Adequacy

Rank	*Parameter*	*Definition*
1	None	The translated sentence does not have meaning
2	Little	The translated sentence conveys some meaning
3	Most	The translated sentence conveys all the meaning of the source sentence
4	All	The translated sentence expresses all the meaning of the source sentence

The geometric average of the parameters was taken and the results are shown in Table 5.

Table 5: Human Evaluation Result

Experiment Type	*Geometric Mean*	
	Fluency	*Adequacy*
Baseline_noAttention	1.925	1.606
Attention_Bahdanau	2.084	2.030
Attention_GNMT	1.953	1.797

Like automatic evaluation result, the human evaluation shows that incorporating attention mechanism in NMT model performs better than the non-attentional models. But, comparing *Experiments II* and *III* (which are both attentional models with different architecture), *Experiment II* exceeds by 0.131 in Fluency and 0.233 in Adequacy parameters. Hence, we can conclude that the performance of both attentional models (*Experiments I* and *II*) are comparable to each other.

Long Distance Dependency

To test the performance of each of the three models in handling long sentence translation, we divided the test data into five groups based on their length using Perl script. Then we made the number of sentence pairs the same for each group. Hence, from the total test data, 100 sentence pairs are used for this experiment. These test data are divided into five groups having each 20 sentence pairs. Then, BLEU scores are computed per group for each model with respect to the length of sentences and we got the result shown in Figure 2.

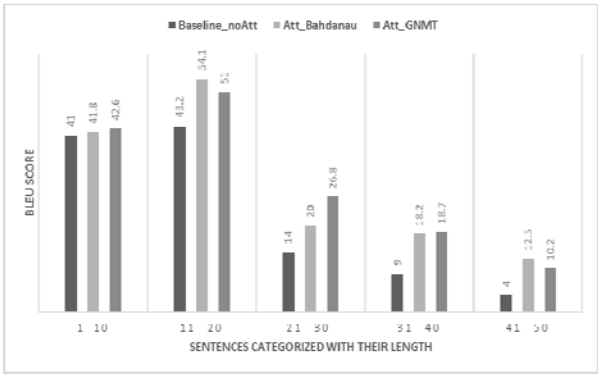

Figure 2: BLEU score on test data with respect to the lengths of sentences

The result as shown in Figure 2 indicates that the performance of Baseline_noAttention model drops deeply as the length of the sentences increases. Both Attention_Bahdanau and Attention_GNMT models perform comparably good with respect to the length of the sentences when compared with the Baseline_noAttentional model.

5. Conclusion and Future Works

This paper mainly focuses on exploring the development of deep learning neural machine model for translating Amharic to English using Recurrent Neural Network. It involves evaluating the effect of NMT in translating from under resourced language Amharic to English with relatively small parallel corpus. The paper has also focused on how far the different NMT models can perform in handling long sentences.

From the experimental results, the following two points are observed:

(1) NMT models that use attention mechanisms are more effective than the baseline non-attentional model. This indicates that incorporating attentional mechanisms in NMT model helps to achieve a high translation performance than with non-attentional models.

(2) In addition, the attentional models outperform the baseline non-attentional model in translating long sentences.

In this paper, we limited the hyper-parameter settings for training the neural network because of limitation of high performance computing hardware resources. Therefore, future work shall focus to

scale-up training a neural network both in terms of computation and memory to achieve a better result.

References

[1] Michael Gasser, “Toward a Rule-Based System for English-Amharic Translation”, In proceedings of the 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages and the 4th Workshop on African Language Technology, May 22, 2012.

[2] Mulu Gebreegziabher and Laurent Besacier, “Preliminary Experiments on English-Amharic Statistical Machine Translation”, In: Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU), Cape town, May 9, 2012.

[3] Mulu Gebreegziabher Teshome, Laurent Besacier, Girma Taye, and Dereje Teferi, "Phoneme-based English-Amharic Statistical Machine Translation", In AFRICON, 2015, pp. 1–5, IEEE.

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

[5] Amjad Almahairi, Kyunghyun Cho, Nizar Habash, and Aaron Courville, "First Result on Arabic Neural Machine Translation", 2016.

[6] Kelleher, John D., “Fundamentals of Machine Learning for Neural Machine Translation”. Translating Europe Forum 2016: Focusing on Translation Technologies, 2016.

[7] Sutskever, I., Vinyals, O., and Le, Q., "Sequence to sequence learning with neural networks", In Advances in Neural Information Processing Systems (NIPS 2014), 2014.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units", In Proceedings of ACL2016, pp. 1715–1725, 2016.

[9] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate”, 2014.

[10] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, and Klaus Macherey, "Google’s neural machine translation system: Bridging the gap between human and machine translation", 2016.

[11] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, "Effective approaches to attention-based neural machine translation", In Proceedings of EMNLP2015, Lisbon, Portugal, September 2015.

[12] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: a simple way to prevent neural networks from over fitting”, Journal of Machine Learning Research, 15(1):1929–1958, 2014

[13] John S. White, Theresa A. O'Connell, and Lynn M. Carlson. "Evaluation of machine translation." In Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics, pp. 206-210, 1993.