HILIGAYNON – CEBUANO SENTENCE TRANSLATOR USING RECURRENT NEURAL NETWORKS

Sean Michael A. Cadigal
University of San Carlos
Cebu City, 6000, Philippines
cadigalsean1219@gmail.com

Christian P. Gelbolingo
University of San Carlos
Cebu City, 6000, Philippines
christiangelbolingo@gmail.com

Christine F. Peña
University of San Carlos
Cebu City, 6000, Philippines
chetpena@yahoo.com

Anthonette D. Cantara
University of San Carlos
Cebu City, 6000, Philippines
toncanan@gmail.com

ABSTRACT

In the Philippines, two of the most spoken dialects are Cebuano and Hiligaynon. Moreover, Cebuano and Hiligaynon are closely related in the sense that both languages share words but have different meaning according to context. The study aimed to create a translation model for Hiligaynon to Cebuano by applying recurrent artificial neural networks with long short-term memory. Two neural networks were developed, one for encoding source sentences and one for decoding target sentences in a manner following sequence-to-sequence learning. The highest accuracy achieved was when the model was trained at 150 epochs and 4000 sentences, yielding a BLEU score of 0.265579454. It can be concluded that a neural machine translation model can be created given sufficient training data.

KEYWORDS

Machine Translation, Natural Language Processing, Cebuano-Hiligaynon, Long Short-Term Memory, Recurrent Neural Networks

1 INTRODUCTION

With the diverse ethnicity present in the Philippines, it is unavoidable to encounter language barriers among the numerous dialects. Two of the most widely used dialects are Cebuano and Hiligaynon. About 9.2% of the Philippine population are native speakers of Hiligaynon while around 20.3% of the population speak Cebuano. Multiple studies have surfaced within the past five years that tackle translation between Philippine dialects. These studies use two neural networks, one as the encoder and another for the decoder. Yaser Al-Onaizan [8] used statistical machine translation to translate between English and Czech. Greenstein and Penner [13] employed a recurrent neural network in translating between English and Japanese. The latter created two neural networks, one to act as the encoder and another to serve as the decoder.

More recent studies conducted by Google were laid on a larger scale as their neural machine translation system was able to handle morphologically rich languages. The following study has presented a neural machine translation system implementing recurrent neural networks with long short-term memory. This system aimed to translate from Hiligaynon to Cebuano. This approach was deemed most appropriate for its ability to adapt its weights from previous outputs. Furthermore, corpora will be constructed from biblical texts from each language.

2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

2.1 Artificial Neural Networks (ANN)

An artificial neural network is a field of machine learning that is modeled after the neurons in the human brain. It is designed in this way so that the system may learn in a similar way as humans do [1]. The significance of implementing artificial neural networks is that it gives the system the ability to learn by giving it correct examples. ANNs have numerous applications in the current world such as pattern recognition, image processing, character recognition and machine translation.

A recurrent neural network differs from the usual feedforward neural network, as the input of a neuron may be its own output from a past calculation. Recurrent networks have possibility to form short-term memory, so they can better deal with position invariance. In recurrent networks, history is represented by neurons with recurrent connections [2]. Recurrent neural networks are commonly used in the fields of natural language processing, handwriting recognition and speech recognition.

Long Short-term memory (LSTM) architecture described by Hochreiter and Schmidhuber [3] is known as a traditional LSTM with forget gates. A "forget gate" controls the extent to which a value remains in memory. And, an "output gate" controls the extent to which the value in memory is used to compute the output activation of the block.

2.2 ENCODER-DECODER TRANSLATOR

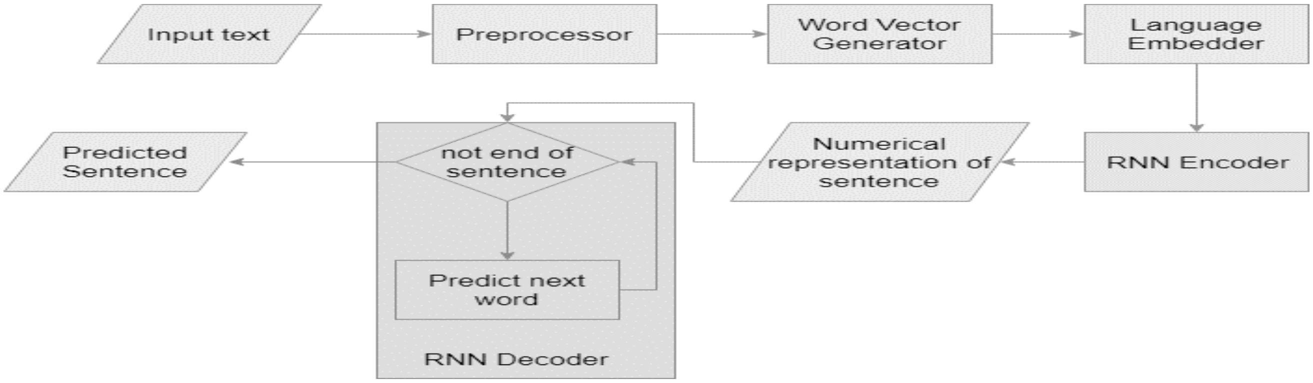

Figure 1. System Flow of RNN Sentence Translator

Figure 1 shows the system flow on how data is being processed in the translator. It accepts an input text in Hiligaynon and outputs the translated Cebuano text.

Preprocessor handles all necessary preprocessing, cleaning of data, changing text to lowercase, removing of special characters excluding hyphen and trimming white spaces. Preprocessed text is then parsed into words and the Word Vector Generator changes each word into its numerical matrix representation known as word vectors that represent the characteristics of a word in a space. Language Embedder repeatedly accepts a word vector until the end of sentence is encountered thus creating a set of word vectors that represents each word in the source sentence. This set of word vectors is used as input by the RNN Encoder to compute the numerical representation of the sentence. The numerical representation is transformed into a vector that is computed by the RNN Decoder to decode or extract information. It outputs the predicted next word until the end of sentence is encountered.

RNN Encoder has 3 layers: the input layer, the processing layer and the output layer. The input layer consists of *n* number of neurons which is equivalent to the number of Hiligaynon words known by the system. The processing layer calculates the state of the Encoder given the word vector of the given word. It is then processed by the last layer of the Encoder which will generate a single numerical representation of the sentence as shown in Figure 2.

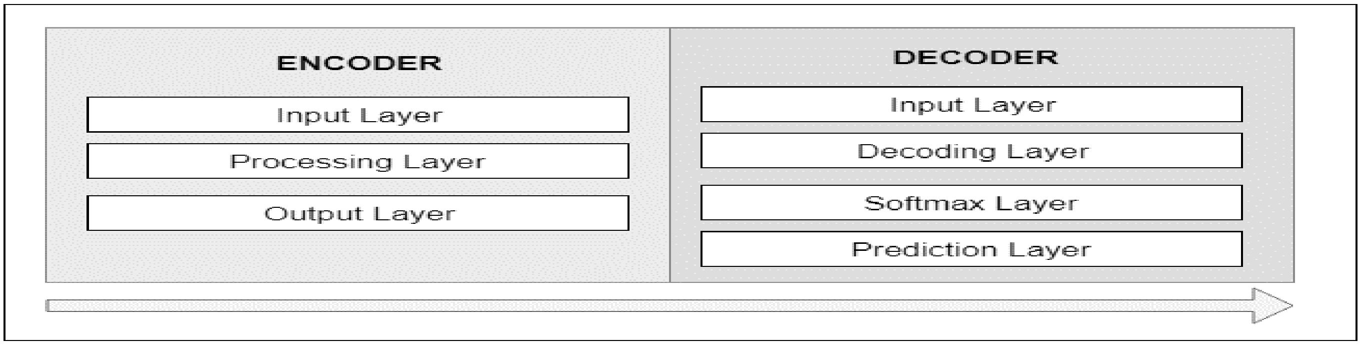

Figure 2: Sentence Translator Layers

Figure 2 also shows that in the Sentence translator, RNN Decoder has 4 layers. Input layer in the RNN Decoder has the same functionality in Encoder; however, the composition of neurons is different. The number of neurons found in the input layer of RNN Decoder is equivalent to the longest sentence in the Cebuano text. Decoding layer decodes the necessary information of the targeted word and then passes the computed data to the next layer, the Softmax layer. It calculates the computed data using the formula:

$$P(y = j\ |x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}}$$

The Softmax layer pinpoints which word will be predicted or guessed and outputs a vector of dimension 1 by the number of Cebuano words known by the system. This output is used by the Prediction layer in order to translate the vector into the predicted word.

2.3 ACCURACY COMPUTATION: BLEU

Bilingual Evaluation Understudy or BLEU is an algorithm for evaluating the quality of text which has been Machine-translated from one language to another. BLEU has been used as a metric since it correlates well with human judgement when comparing the output of machine translation systems [4]. BLEU uses modified n-gram precision. BLEU's n-gram precision is modified to eliminate repetitions that occur across sentences [5]. BLEU's precision score p_n is calculated for each n-gram length by summing over the matches for every hypothesis sentence S in the complete corpus C as [6]:

$$p_n = \frac{\sum_{S \in C} \sum_{n\ grams \in S} Count_{matched}(n\ gram)}{\sum_{S \in C} \sum_{n\ grams \in S} Count(n\ gram)}$$

Since BLEU is precision based, and recall is difficult to formulate over multiple reference translations, a brevity penalty is introduced to compensate for the possibility of proposing high-precision hypothesis translations which are too short [5]. Brevity Penalty is calculated as:

$$BP = \begin{cases} 1 & if \;\; c > r \\ e^{1-r/c} & if \; c \;\leq r \end{cases}$$

where c is the length of the corpus of hypothesis translations,

r is the effective reference corpus length.

Thus, the BLEU score is calculated as:

$$BLEU = BP * \exp(\sum_{n=1}^{N} W_n \log p_n)$$

3 RESULTS AND DISCUSSION

Table 1. Results on BLEU Score on different number of sentences and epochs

	Number of Sentences			
Epochs	1000	2000	3000	4000
100	0.296736406	0.277305609	0.298439085	0.257849909
150	0.28098048	0.30414673	0.25518989	0.26557945
200	0.24733645	0.29633015	0.26056444	0.26010842

As shown in Table 1, the highest accuracy achieved was at 2000 sentences and 150 epochs. However, if the training data is insufficient, the translator repeats a single phrase or sentence for every sentence it has to translate. This behavior is shown starting at 1000 sentences with 100 epochs and is last observed at 3000 sentences and 100 epochs. The reason that these cases scored higher BLEU scores than rest is that it used the most common words as the translation. At 3000 sentences and 150 epochs up until 4000 sentences and 200 epochs, there is little to none repetition of words and/or sentences between translations. The highest BLEU score with no repetition of words and/or sentences occurs at 4000 sentences and 150 epochs. With insufficient data and training, the translation model resorted to outputting the most commonly used words in the vocabulary. The highest BLEU score achieved with non-repeating translations is 0.265579454 which occurs when the model is trained at 4000 sentences and 150 epochs

A study conducted by Lilt Labs [4] tested different well-known commercially available translators against each other using BLEU score as a metric. These included translators created by Google, Microsoft and Systran. They were tested with English to French and English to German translations. Their BLEU scores ranged from 27% to 35% in English to French translations and 22% and 30% in English to German translations. In comparison to these, the Hiligaynon - Cebuano translator scored fairly close in BLEU score.

4 CONCLUSIONS

It can be concluded that a machine translation model using recurrent neural networks and long short-term memory can be created using parallel corpora as training data as shown in the presented study. The created translation model achieved a peak BLEU score of 0.265579454.

ACKNOWLEDGMENTS

We would like to extend our heartfelt gratitude to the Department of Computer and Information Sciences of the University of San Carlos for the academic support, to Christian Dave Baclayon and Andre Cedric Peñalosa for letting us assist in their study and teaching us the basics of artificial neural networks, to Philogene Kyle Dimpas and Royce Vincent Po for performing their study in parallel to ours and assisting us in learning the concepts of artificial neural networks in addition to helping train our translation model, to Marianne Mithun for helping gather related literature relating to the topics being researched, to our friends and family for their everlasting support, and lastly to almighty God for giving us the opportunities we needed to tackle our project.

REFERENCES

[1] D. S. Christos Stergiou, "Neural Networks," [Online]. Available: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

[2] T. Mikolov, Recurrent Neural Network Based Language Model, Brno University of Technology, Johns Hopkins University, 2010.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, pp. 1735-1780, 1997.

[4] "Lilt Labs," 2 August 2017. [Online]. Available: https://labs.lilt.com/2017-machine-translation-quality-evaluation-603ff3ec3c36.

[5] C. O. M. a. K. P. Callison-Burch, "Re-evaluating the Role of BLEU in Machine Translation Research," 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL, p. 249–256, 2006.

[6] S. R. T. W. a. W. Kishore Papineni, "Bleu: A method for automatic evaluation of machine translation," in ACL, 2002.

[7] B. K. S. Baljinder Kaur, "Machine Translation: An Analytical Study," Baljinder Kaur et al Int. Journal of Engineering Research and Applications, pp. 168-175, 2014.

[8] Yaser Al-Onaizan, "Statistical Machine Translation," 1999.

[9] Kyunghyun Cho, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," 2014.

[10] K. Y. Dzmitry Bahdanau, "Neural Machine Translation By Jointly Learning to Align and Translate," 2016.

[11] R. Dale, H. Moisl and H. Somers, Handbook of Natural Language Processing, New York: Marcel Dekker, 2000.

[12] L. Schubert, The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2015.

[13] E. Greenstein and D. Penner, Japanese-to-English Machine Translation Using Recurrent Neural Networks, Stanford University, 2015