IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 11, No. 2, June 2022, pp. 658~665
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i2.pp658-665

Sequence-to-sequence neural machine translation for English-Malay

Yeong-Tsann Phua, Sujata Navaratnam, Chon-Moy Kang, Wai-Seong Chew
School of Computing and Creative Media, UOW Malaysia KDU University College, Selangor, Malaysia

Article Info

Article history:

Received Dec 29, 2020
Revised Jan 20, 2022
Accepted Feb 5, 2022

Keywords:

Encoder-decoder
Machine translation
Malay language
Neural machine translation
Sequence-to-sequence

ABSTRACT

Machine translation aims to translate text from a specific language into another language using computer software. In this work, we performed neural machine translation with attention implementation on English-Malay parallel corpus. We attempt to improve the model performance by rectified linear unit (ReLU) attention alignment. Different sequence-to-sequence models were trained. These models include long-short term memory (LSTM), gated recurrent unit (GRU), bidirectional LSTM (Bi-LSTM) and bidirectional GRU (Bi-GRU). In the experiment, both bidirectional models, Bi-LSTM and Bi-GRU yield a converge of below 30 epochs. Our study shows that the ReLU attention alignment improves the bilingual evaluation understudy (BLEU) translation score between score 0.26 and 1.12 across all the models as compare to the original Tanh models.

Corresponding Author:

Yeong-Tsann Phua
School of Computing and Creative Media, UOW Malaysia KDU University College
Utropolis Glenmarie, Jalan Kontraktor U1/14, Seksyen U1, 40150 Shah Alam, Selangor, Malaysia
Email: yt.phua@kdu.edu.my

1. INTRODUCTION

Malay language is part of the Nusantara in Austronesia language family [1]. This language is spoken by 290 million people across the world. This language is the national language of Malaysia and it is widely used in both public and private sectors in the country. In Malaysia, this language adopted Roman alphabet during the British administration period [2]. The Malaysian government is actively promoting the country to be a hub for education and medical tourism. In 2018, there were over 127,000 foreign students in Malaysia. The number reached 130,000 in 2019. On the other hand, over one million medical tourists arrived at Malaysia in year 2017. The number reached 1.3 million in year 2020. The need of suitable machine translation (MT) is essential to help international students and tourists to understand conversation and content when dealing with the locals [3].

As a type of natural language process (NLP) application [4], MT involves the process of using computer software to translate messages from a specific language into another language [5], [6], [7]. This process involves a source natural language (e.g., English) and a target natural language (e.g.: Malay) [6]. This is the essential process [8] in news translation, movie subtitling, question/answer systems and chatbots with understanding of different languages [9]. Two common state-of-the art approaches [5] are statistical machine translation (SMT) and neural machine translation (NMT) [10], [11].

The statistical machine translation takes the word-to-word approach between the source and target words. The process involves statistical analysis using the text corpora [12]. Further enhancement of the approach will restrict the alignment of each source word with exactly one target word [13], [14]. The similar approach is used in speech recognition by applying hidden markov model (HMM).

On the other hand, NMT adopted deep neural network [15] using recurrent neural network (RNN), long short-term memory (LSTM) and gated recurrent unit (GRU). The fundamental unit in NMT is a vector

[16]. NMT depends on a word embedding to transform the word sequence into a vector before the model training can take place [17]. Besides that, there are some work done by combining both SMT and NMT to take advantage of the strength both models [5]. Some of these works include Stahlberg *et al*. that used risk estimation in NMT [18] and Du and Way's cascade framework in the hybrid MT [19].

All the MTs require parallel text corpus to train the models. The preparation of a parallel text corpus is an intensive data-driven process. The SMT will require additional corpus of the target language to formulate the language model. Traditionally, SMT will perform well in small datasets with long sentences [20]. This approach demonstrated better performance compared to NMT with a domain mismatch between training and testing datasets.

MT from English to other languages had been introduced more than 35 years [21]. But, the study of Malay language in MT begun around 1984 by Unit Terjemahan Melalui Komputer (UTMK) at Universiti Sains Malaysia [22]. The first online English-Malay MT system was introduced in 2002 through the collaboration between MIMOS and USM which was aimed at the translation gist [23]. Later in 2006, example-based machine translation (EBMT) uses bilingual corpus examples to form proper representation for the translation [23]. Google Translate is another popular platform for MT [21], [24]. In terms of NMT for English-Malay MT, there is very little research was carried out.

In this manuscript, a rectified linear unit (ReLU) based attention score has been proposed to improve the performance of RNN-based NMT on conversational dialogue in English-Malay translation. Intuitively, this enhanced attention-based sequence-to-sequence NMT will be able to preserve the long sequence context vector and prevent common vanishing gradient problem in the deep networks. In this paper, section 2 will consist of a brief overview of sequence-to-sequence model. Section 3 will discuss the experiment setup. Section 4 will discuss the result and performance of various models used in the experiment.

2. RELATED WORKS

The recurrent neural network (RNN) consists of recurrent cells which the current state of the cell depends on both past cell states and existing input in feedback connection. The RNN unit suffers two major problems, the exploding gradients, and vanishing gradients [25]. This is due to the weakness of RNN unit that cannot handle long-term dependencies. In this experiment, the RNN-based sequence-to-sequence (Seq2seq) NMT models were used to compare their performance. These RNN models are: i) long short-term memory (LSTM), ii) bidirectional LSTM (Bi-LSTM), and iii) gated recurrent unit (GRU).

2.1. Long short-term memory (LSTM)

The long short-term memory (LSTM) was proposed by Hochreiter and Schmidhuber [26]. This RNN based neural netwell uses gates to retain information in the cell. This architecture is capable to deal with the long-term dependencies issue suffers in RNN. There are three gates in LSTM, the input gate, forget gate and output gate. The input gate takes in previous hidden state and current input. It decides which values will be updated with a sigmoid function. The forget gate decides which information from previous hidden state and current input to retain or discard. Lastly, the output gate decides what the next hidden state should be.

2.2. Bidirectional LSTM (Bi-LSTM)

The main idea behind Bi-LSTM is to combine input information in the past and future of a specific time step in LSTM model [27]. This architecture facilitates more input information in the network by allowing the network to preserve past future information. The implementation consists of a regular RNN unit that has two directions or states, one for positive time direction or called forward states and another direction in negative time called backward states.

2.3. Gated recurrent unit (GRU)

Gated recurrent unit simplifies the LSTM network by removing the cell state in the network. It uses a hidden state to transfer information. There are only two gates, the reset and update gates in GRU [28], which have the advantage of retaining information from long ago. The update gate will determine the amount of information from the past time step to pass along to the future. Meanwhile, the reset gate will decide the amount of past information to retain.

2.4. Sequence-to-sequence (seq2seq)

In the original sequence-to-sequence (seq2seq) model introduced by Sutskever *et al*. [29], it has two major components, an encoder, and a decoder [29]. The encoder consists of a stack of recurrent units where it will take in each element in the input sequence. It will collect information about its internal state to form internal state vector or called content vector. Then, it will forward it through propagation. The hidden state h_i is computed by (1) using the existing input x_t, previous state h_{t-1} and the network weight, W.

$$h_t = f(W^{hh}h_{t-1} + W^{hx}x_t) \tag{1}$$

At the other end, the decoder also consists of a stack of recurrent units where it will predict an output at each time step t. The initial state of the decoder is initialized from the final states of the encoder. Each of the recurrent unit will accept a hidden state from the previous unit and compute its own hidden state. The hidden state h_i of the decoder is computed using (2).

$$h_t = f(W^{hh}h_{t-1}) \tag{2}$$

Then, the output y_t at time step t is computed using (3). This requires the combination of both hidden state of the existing time step and respective weight W^S. The Softmax function is applied to generate the probability vector of output.

$$y_t = softmax(W^S h_t) \tag{3}$$

The result achieved in Sutskever *et al*. [29] model is 34.81 in BLEU score which is above the SMT baseline which is 33.30.

2.5. Attention mechanism

The attention mechanism was first introduced in Bahdanau *et al*. [10]. It aims to solve representation issue in seq2seq model. In seq2seq, the decoder only received the last encoder's hidden state. The attention mechanism works as part of the network to capture the important parts of the source [30]. This mechanism works an interface between the encoder and decoder. Hence, the decoder is provided with all the encoder's hidden states [31].

The seq2seq model with attention implementation consists of the encoder, decoder, and attention layers. Within the attention layer, there are three components which include alignment layer, attention weights and context vector. The alignment layer maps the input at time step t and the output from previous time step $t-1$. This is based on the previous state h_{t-1} and previous state s_{p-1}. The alignment score is

$$r_{rp} = v_a^T tanh(W^{ss}s_{p-1} + W^{hh}h_{t-1}) \tag{4}$$

In this experiment, the hyperbolic tangent, tanh function is replaced with ReLU function. Hence, equation (4) will become,

$$r_{rp} = v_a^T ReLU(W^{ss}s_{p-1} + W^{hh}h_{t-1}) \tag{5}$$

This adjustment aims to enhance the alignment score to overcome the common vanishing gradient issue which commonly occurs in tanh alignment score [32], [33].

The alignment score is computed using (6).

$$\alpha_{tp} = \frac{\exp(r_{rp})}{\sum_{t=1}^{|x|} \exp(r_{rp})} \tag{6}$$

The context vector c_p requires the previous state h_{t-1}, previous state s_{p-1} and alignment score as shown in (7).

$$c_p = \sum_{t=1}^{|x|} \alpha_{tp} h_t \tag{7}$$

Hence, the decoder will generate output with next target hidden state by accepting input from previous state y_{p-1} and source context vector c_p as shown in (8).

$$s_p = f(W^{ss}s_{p-1} + W^{sy}y_{p-1} + W^{sc}c_p) \tag{8}$$

The j^{th} decoder's target hidden state requires the previous hidden state as in (9).

$$t_j = f(W^{ss}s_j + W^{sy}y_{j-1} + W^{sc}c_j) \quad (9)$$

Finally, output word is produced using the probability distribution P_j using the Softmax function using (10).

$$P_j = softmax(W^s t_j) \quad (10)$$

3. METHODOLOGY

3.1. The English-Malay parallel text corpus

In this experiment, the English-Malay parallel corpus were collected. The compiled corpus consists of parallel text for models training and test purpose. These parallel texts were extracted from the following sources: i) bilingual sentence pairs from ManyThings.org.; ii) local Malay movie bilingual subtitles; and iii) translated English-Malay bilingual translation corpus [34].

All these corpuses are not in ready form. Hence, some pre-processing was required to compile it into single bilingual sentence pairs corpus [6]. In this study, the pre-processing is required allow better processing for the algoritm [35]. These processing involves:

– Data loading: This step involves loading all the data from different sources, comma delimited format (csv) text files and JSON format into single csv file.
– Lowercasing: This step converts the text to lowercase form to prevent variation in mixed case typing in text and sparsity issue.
– Punctuation, symbols removal and non-text character removal: All the non-text characters in the data are removed to allow the language model fully trained on text-based tokens.
– Word tokenization: This step involves splitting the text into word token before feeding into the model for training.

3.2. Evaluation

The bilingual evaluation understudy (BLEU) score was used in this experiment to evaluate the quality of the translation. This score compared the translated text with the original reference translation text [36]. The evaluation involves matching n-grams in the target translation with the n-grams reference text. This evaluation matrix has these advantages: i) it is quick and simple to calculate, ii) it is language independent, iii) it has high correlation with human evaluation, and iv) it is widely adopted the NMT for evaluation.

In this experiment, four models were trained, and the models' BLEU scores were computed. These models are: i) vanilla LSTM seq2seq, ii) LSTM seq2seq with attention mechanism using tanh alignment and ReLU alignment, iii) GRU seq2seq with attention mechanism using tanh alignment and ReLU alignment, iv) bidirectional-LSTM seq2seq with attention mechanism using tanh alignment and ReLU alignment, and v) bidirectional-GRU seq2seq with attention mechanism using tanh alignment and ReLU alignment

Early stopping was introduced in the model training. This implementation was introduced to prevent overfitting during training. The mechanism used the training's validation loss to determine when to stop the model training.

4. RESULT AND DISCUSSION

In this experiment, all the models were setup and configured using Google Tensorflow-GPU 2.2. The parallel corpus used for training consists of 189,000 pairs of bilingual English-Malay sentence pairs. The testing dataset consists of 199 pairs of bilingual sentence pairs. Total vocabulary from source and target were 8183 and 6938 word respectively. The out of vocabulary (OOV) token was incorporated to substitute words that did not exist in the embedding. Early stopping was incorporated in the models training. All the models' output was evaluated using BLEU score. Hence, the reference text in the dataset must consist of at least 4 words.

A vanilla LSTM seq2seq model was used as the baseline model. This vanilla LSTM seq2seq model consisted of both encoder and decoder that had a 300-dimension embedding and a single hidden LSTM layer with 512 neurons. During the training, this model stopped at epoch 44. The model achieved a BLEU score of 80.39. The same test dataset was loaded into Lingvanex.com for translation and the score of the translation is 62.04.

Next, four different seq2seq models were setup and trained. These models incorporated with Bahdanau attention mechanism [10]. Table 1 shows the training epoch for all the models. Generally, all models converaged faster when incorporated the attention mechanism in the seq2seq models as compared to the vanilla model. All these models achieved validation loss that are below 0.36 and converged between epoch 24 and 38. Among these models, the bidirectional models such as Bi-LSTM and bidirectional GRU (Bi-GRU) took 24 and 27 epochs or about 39% less epoch to converge in the training.

Table 1. Training epoch and duration for models

Model	No of Epoch	Duration for each epoch
Vanilla LSTM model (baseline model)	44	106s 100ms/step
LSTM Tanh alignment	31	143s 134ms/step
LSTM ReLU alignment	37	144s 135ms/step
GRU Tanh alignment	37	135s 126ms/step
GRU ReLU alignment	38	133s 124ms/step
BiLSTM Tanh alignment	24	247s 232ms/step
BiLSTM ReLU alignment	24	250s 235ms/step
BiGRU Tanh alignment	27	234s 219ms/step
BiGRU ReLU alignment	25	233s 218ms/step

Table 2 shows the samples from the various models. From the experiment, all the models achieved higher BLEU scores between 0.90 and 4.57 as compares to the baseline model. Among the models, Bi-LSTM with ReLU attention mechanism was able to achieve BLEU score of 85.14 which is about 4.75 better than the vanilla model. This followed by Bi-GRU model with ReLU attention mechanism at BLEU score of 83.74 and 3.35 above the vanilla mode. Generally, the models with ReLU attention alignment were able to achieve higher accuracy as compared to the Tanh attention alignment from 0.26 in LSTM model to 1.12 in Bi-LSTM model.

Tables 3 to 6 show the attention weights of translation samples from Bi-LSTM model with Tanh and ReLU attention alignment. Based on the samples, the ReLU attention alignment model generally has higher weights as compared to the Tanh attention alignment. Besides that, the weights are aligned closely to the intended output words. On top of that, the emphasis of the attention weights in ReLU attention alignment are relatively stronger on the input token as compared to other tokens in the sequence.

Table 2. BLEU score for testing result of seq2seq models with attention mechanism

	Attention Aligment	
Model	Tanh	ReLU
LSTM	83.38	83.65
GRU	81.28	81.63
BiLSTM	84.02	**85.14**
BiGRU	83.45	83.74

Table 3. Attention weights for Bi-LSTM with Tanh attention alignment for sample result 1

	perancis	adalah	di	eropah	barat	
france	**9.97E-01**	1.40E-03	5.79E-05	1.24E-04	7.20E-05	1.91E-03
is	9.04E-04	1.27E-01	2.17E-03	1.65E-04	8.54E-04	**1.51E-01**
in	9.02E-04	7.86E-01	**8.44E-01**	4.64E-03	1.48E-03	1.36E-01
western	4.72E-04	2.15E-02	1.31E-03	1.01E-02	**9.88E-01**	1.13E-01
europe	2.79E-04	1.33E-02	1.52E-01	**9.84E-01**	4.15E-03	3.61E-02

Table 4. Attention weights for Bi-LSTM with ReLU attention alignment sample result 1

	perancis	adalah	di	eropah	barat	
france	**1.00E+00**	3.55E-03	1.27E-06	5.01E-09	1.65E-07	1.74E-03
is	7.44E-05	**3.49E-01**	2.48E-05	4.97E-08	3.69E-05	2.62E-02
in	4.36E-05	6.30E-01	**9.88E-01**	8.70E-05	2.26E-03	2.25E-02
western	9.51E-08	3.99E-03	1.20E-06	2.14E-04	**9.29E-01**	2.94E-03
europe	1.91E-06	5.88E-03	1.18E-02	**1.00E+00**	3.90E-02	8.85E-03

Table 5. Attention weights for Bi-LSTM with Tanh attention alignment sample result 2

	kami	mempunyai	masa	yang	baik	
we	**8.90E-01**	1.45E-03	2.43E-04	1.34E-03	1.30E-03	4.03E-02
are	5.88E-02	9.35E-03	1.35E-04	1.32E-03	5.36E-04	2.76E-02
having	3.33E-02	**8.70E-01**	1.53E-02	3.08E-02	6.19E-03	1.45E-01
good	2.64E-03	1.53E-03	6.80E-04	**9.41E-01**	**9.78E-01**	8.51E-02
time	5.18E-03	1.14E-01	**9.72E-01**	1.96E-02	1.36E-02	1.49E-01

Table 6. Attention weights for Bi-LSTM with ReLU attention alignment sample result 2

	kami	mempunyai	masa	yang	baik	
we	**8.19E-01**	1.34E-04	5.19E-08	6.24E-05	3.54E-06	3.35E-04
are	6.48E-02	1.34E-03	5.67E-06	2.67E-04	4.77E-06	3.80E-04
having	6.59E-02	**9.95E-01**	8.52E-02	2.22E-01	1.20E-03	2.78E-02
good	1.16E-03	3.40E-06	2.40E-07	**7.07E-01**	**9.99E-01**	1.38E-03
time	5.84E-03	3.53E-03	**9.15E-01**	6.46E-02	1.12E-04	4.23E-02

5. CONCLUSION

In this paper, we empirically evaluated different seq2seq models based on the attention alignment for neural machine translation in English to Malay language. The evaluation focused on task of sequence modelling using English-Malay bilingual parallel text corpus. As there is very limited work done using neural machine translation in this area, this paper focuses on the used of ReLU attention alignment to improve the performance of the translation. Generally, the Bi-LSTM and Bi-GRU are able to achieve higher BLEU score as compared to the original Tanh alignment score which as confirmed by the results.

ACKNOWLEDGEMENT

The research is fully sponsored by UOW Malaysia KDU Grant (KDURG/2018/1/002). Special thanks to staffs who have contributed either directly or indirectly to the research.

REFERENCES

[1] N. S. Karim, F. M. Onn, H. H. Musa, and A. H. Mahmood, *Tatabahasa Dewan Edisi Ketiga*, 3rd ed. Dewan Bahasa Dan Pustaka, 2015.

[2] O. Sulaiman, *Malay for Everyone*. Petaling Jaya, Malaysia: Pelanduk Publications (M) Sdn. Bhd., 2019.

[3] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," *Mach. Transl.*, vol. 32, no. 1–2, pp. 167–189, 2018, doi: 10.1007/s10590-017-9203-5.

[4] N. G. Kharate and V. H. Patil, "Inflection rules for Marathi to English in rule based machine translation," *IAES Int. J. Artif. Intell.*, vol. 10, no. 3, p. 780, 2021, doi: 10.11591/ijai.v10.i3.pp780-788.

[5] Y. L. Yeong, T. P. Tan, K. H. Gan, and S. K. Mohammad, "Hybrid machine translation with multi-source encoder-decoder long short-term memory in English-Malay translation," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4–2, pp. 1446–1452, 2018, doi: 10.18517/ijaseit.8.4-2.6816.

[6] B. N. V. Narasimha Raju, M. S. V. S. Bhadri Raju, and K. V. V. Satyanarayana, "Effective preprocessing based neural machine translation for english to telugu cross-language information retrieval," *IAES Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 306–315, 2021, doi: 10.11591/ijai.v10.i2.pp306-315.

[7] Y. Jia, M. Carl, and X. Wang, "How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study," *J. Spec. Transl.*, no. 31, pp. 60–86, 2019.

[8] S. Aneja, S. Nur Afikah Bte Abdul Mazid, and N. Aneja, "Neural Machine Translation model for University Email Application," *ACM Int. Conf. Proceeding Ser.*, vol. 18, pp. 74–79, 2020, doi: 10.1145/3421515.3421522.

[9] J. Yu, Z. J. Zha, and J. Yin, "Inferential machine comprehension: Answering questions by recursively deducing the evidence chain from text," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 2241–2251, 2020, doi: 10.18653/v1/p19-1217.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ICLR 2015*, Sep. 2015.

[11] Y. Z. Low, L. K. Soon, and S. Sapai, "A Neural Machine Translation Approach for Translating Malay Parliament Hansard to English Text," *2020 Int. Conf. Asian Lang. Process. IALP 2020*, pp. 316–320, 2020, doi: 10.1109/IALP51396.2020.9310470.

[12] H. Sujaini, "Improving the role of language model in statistical machine translation (Indonesian-Javanese)," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 2102–2109, 2020, doi: 10.11591/ijece.v10i2.pp2102-2109.

[13] F. J. Och, C. Tillmann, and H. Ney, "Improved Alignment Models for Statistical Machine Translation," in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[14] C. Gulcehre, O. Firat, K. Xu, K. Cho, and Y. Bengio, "On integrating a language model into neural machine translation," *Comput. Speech Lang.*, vol. 45, pp. 137–148, 2017, doi: 10.1016/j.csl.2017.01.014.

[15] O. Firat, K. Cho, B. Sankaran, F. T. Yarman Vural, and Y. Bengio, "Multi-way, multilingual neural machine translation," *Comput. Speech Lang.*, vol. 45, pp. 236–252, 2017, doi: 10.1016/j.csl.2016.10.006.

[16] P. Basmatkar, H. Holani, and S. Kaushal, "Survey on neural machine translation for multilingual translation system," *Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019*, no. Iccmc, pp. 443–448, 2019, doi: 10.1109/ICCMC.2019.8819788.

[17] M. Maimaiti, Y. Liu, H. Luan, and M. Sun, "Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 150–163, 2022, doi: 10.26599/TST.2020.9010029.

[18] F. Stahlberg, A. De Gispert, E. Hasler, and B. Byrne, "Neural machine translation by minimising the bayes-risk with respect to syntactic translation lattices," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, no. 2008, pp. 362–368, 2017, doi: 10.18653/v1/e17-2058.

[19] J. Du and A. Way, "Neural Pre-Translation for Hybrid Machine Translation," in *MT Summit XVI*, 2017, vol. 1, pp. 27–40.

[20] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," in *Proceedings of the First Workshop on Neural Machine Translation*, 2017, pp. 28–39, doi: 10.18653/v1/W17-3204.

[21] N. Yusoff, Z. Jamaludin, and M. H. Yusoff, "Semantic-based Malay-English translation using n-gram model," *J. Telecommun. Electron. Comput. Eng.*, vol. 8, no. 10, pp. 117–123, 2016.

[22] K. A. Rahman and N. M. Norwawi, "Proverb Treatment in Malay-English Machine Translation," *2nd Int. Conf. Mach. Learn. Comput. Sci.*, pp. 4–8, 2013.

[23] A. R. Suhaimi, A. Noorhayati, H. Hafizullah Amin, and D. Abdul Wahab, "Real time on-line english-malay machine translation (MT) system," *Third Real-Time Technol. Appl. Symp.*, pp. 229–239, 2006, doi: 10.1.1.126/6692.
[24] M. Yamada, "The impact of google neural machine translation on post-editing by student translators," *J. Spec. Transl.*, no. 31, pp. 87–106, 2019.
[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 4, pp. 2047–2052, doi: 10.1109/IJCNN.2005.1556215.
[26] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[27] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.
[28] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *Math. Program.*, vol. 155, no. 1–2, pp. 105–145, Sep. 2014, doi: 10.1007/s10107-014-0839-0.
[30] A. Vaswani *et al.*, "Attention Is All You Need," *NIPS'17 Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, no. Nips, pp. 6000–6010, Jun. 2017.
[31] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018, doi: 10.1016/j.neucom.2018.01.007.
[32] Y. Yu, K. Adu, N. Tashi, P. Anokye, X. Wang, and M. A. Ayidzoe, "RMAF: Relu-Memristor-Like Activation Function for Deep Learning," *IEEE Access*, vol. 8, pp. 72727–72741, 2020, doi: 10.1109/ACCESS.2020.2987829.
[33] T. Szandała, "Review and comparison of commonly used activation functions for deep neural networks," *Stud. Comput. Intell.*, vol. 903, pp. 203–224, 2021, doi: 10.1007/978-981-15-5495-7_11.
[34] Z. Husein, "Malay-Dataset," *GitHub repository*, 2018. [Online]. Available: https://github.com/huseinzol05/Malay-Dataset.
[35] O. J. Ying, M. M. A. Zabidi, N. Ramli, and U. U. Sheikh, "Sentiment analysis of informal malay tweets with deep learning," *IAES Int. J. Artif. Intell.*, vol. 9, no. 2, pp. 212–220, 2020, doi: 10.11591/ijai.v9.i2.pp212-220.
[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, vol. 371, no. 23, p. 311, doi: 10.3115/1073083.1073135.

BIOGRAPHIES OF AUTHORS

Yeong-Tsann Phua received his B. Sc. in Computer with Education from Universiti Teknologi Malaysia (UTM) in 1997 and his Master of Computer Science from Universiti Malaya (UM) in 2006. His research interests include Machine Learning, Deep Learning, Natural Language Processing and Computer Vision. His current research is in text summarization in Malay language. Phua is currently a Ph. D student in Universiti Teknologi PETRONAS (UTP). Email: yt.phua@kdu.edu.my.

Sujata Navaratnam Sujata holds a Bachelor (Hons) in Computer Science from Universiti Sains Malaysia (USM) and a Masters in Software Engineering from Open University Malaysia (OUM). She is currently pursuing her PhD in Science and Technology with Universiti Sains Islam Malaysia (USIM) where she is focusing on computer security and data analytics. Email: sujata.n@kdu.edu.my.

Kang Chon Moy received her B. Sc. (Hons) of Business Information Technology from University of Central England, Birmingham in 1998 and her M. Sc. in Business Information Technology, from RMIT University, Australia in 2003. Her research interests include cloud computing, e-learning, security and networking and machine mearning. Her current research is in natural language processing in Malay language. Email: cm.kang@kdu.edu.my.

Wai-Seong Chew received her B. Sc. of Science (Majoring in Physics) from University Putra Malaysia (UPM) in 1999 and her M. Sc. in Information Technology from University Putra Malaysia (UPM) in 2001. Her research interests include Machine Learning, Big Data and Computer Vision. Chew is currently working in UOWMKDU University College as a Program Leader. Email: wschew@kdu.edu.my.