Malay phoneme-based subword news headline generator for low-resource language

Yeong Tsann Phua[1,2], Kwang Hooi Yew[1], Mohd Fadzil Hassan[1], Matthew Teow Yok Wooi[3]
[1]Department of Computer and Information Science, Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia
[2]School of Diploma Studies, Taylor's College, Subang Jaya, Malaysia
[3]University Partnership (Coventry University), PSB Academy, Raffles, Singapore

Article Info

Article history:

Received Dec 18, 2023
Revised May 17, 2024
Accepted Jun 1, 2024

Keywords:

Low-resource language
Malay language
Natural language processing
News headline generator
Subword embedding

ABSTRACT

The booming of technology has significantly increased the amount of news articles for readers. The headline of news plays an essential role in attracting readers. Traditionally, crafting the news headline is a manual task at the news desk. The motivation of this paper is to address the issues faced in low-resource languages, such as the Malay language. The main contribution of this paper is a new hybrid model based on extractive- and abstractive-based text summarization with the integration of a geographical linguistics model; a Malay phoneme-based subword embedding has been developed to solve the complex morphological issue in the Malay language-based computational linguistic applications. The experiment involves various sequence-to-sequence (seq2seq) models to generate the Malay news headlines. Besides that, the out-of-vocabulary (OOV) is assessed in the models. From the experiment, the proposed hybrid text summarization model shows significant improvement over the baseline models above 11.00 in ROUGE-1, 4.00 ROUGE-2, and 11.00 in ROUGE-L. The proposed model can reduce the OOV rate to below 15%.

Corresponding Author:

Yeong Tsann Phua
Department of Computer and Information Science, Universiti Teknologi PETRONAS
Seri Iskandar, Perak Darul Ridzuan 32610, Malaysia
Email: yeong_17008256@utp.edu.my

1. INTRODUCTION

In the world of journalism, news headlines serve as the agents to provide the essence of news articles [1]. The headline plays a pivotal role in providing the main point of the news story while attracting and motivating readers to explore the news content further. Moreover, a well-crafted headline will reduce the reader's cognitive burden while reading the news [2]. Therefore, constructing efficient and accurate news headlines is a crucial journalistic skill because it shapes how the news is consumed, shared, and perceived by the audience. However, in the current fast-paced news environment, journalists and editors often face time constraints. Under these circumstances, automatic news headline generation (NHG) systems help to streamline the process and save time as well as resources by quickly producing headlines based on the content of the news story.

The NHG is considered a type of text summarization. Text summarization can occur for both single or multiple documents [3]. Generally, there are two approaches adopted in NHG, the extractive and the abstractive approaches [4]. Both approaches aim to produce short yet accurate headlines that convey the main idea of the news content. The extractive approach aims at several important sentences and applies sentence compression techniques to produce the headline [4]. Due to the reuse of content words in the headline, distortion of meaning will occur in the headline produced [5]. The abstractive approach applies advanced natural language processing (NLP) techniques to produce the headline. Some of these techniques allow paraphrasing, synonymous substitutions, and sentence contractions [6]. The neural network-based attention mechanism encoder-decoder models [7] have demonstrated impressive results such as in [8], [9].

As NHG becomes an important task in NLP [6], research on and development of the systems mainly focused on commonly spoken languages such as English, while overlooking low resource languages (LRL). Most of the state-of-the-art approaches in NHG were able to produce impressive results with training using large datasets such as Gigaword, which is in English [10], [11]. In contrast, LRL, like many Asian languages including the Malay language, are often overlooked because they lack linguistic resources such as part of speech (POS) taggers and corpora [12] The limited availability of LRL dataset has led to challenges in the development of automatic headline generation specifically tailored for LRL. This situation is further complicated by the Malay language’s rich morphological structure, which often results in a high out-of-vocabulary (OOV) rate when traditional word embedding methods are employed. To navigate these obstacles, this study introduces a Malay salient sentence extractor that leverages sentence extraction methods to distill essential content, thereby streamlining both the model training process and the headline generation task.

In response to the limitations imposed by the availability of resources and the intrinsic language features, the authors proposed a sequence-to-sequence (seq2seq) deep learning-based model named Malay salient abstractive summarizer (MSAS). It is believed that innovative approaches to NHG for the Malay language can offer promising solutions. This research adopts a hybrid seq2seq approach of extractive and abstractive methods tailored to the Malay linguistic context. The Malay phoneme-based subword embedding has been employed to effectively address the pervasive OOV issue [13], substantially reducing the embedding layer’s vocabulary size without compromising the model's linguistic comprehension [14]. Through these methodological advancements, this study not only contributes to the body of knowledge in NLP but also offers practical solutions for automated headline generation in the context of LRL. The experiment is conducted using a dataset of 45,000 Malay language news stories and their headlines.

The related work is described in section 2. Section 3 provides the proposed MSAS model architecture. The experiment setup, execution, and result are discussed in section 4. The conclusion is presented in section 5.

2. RELATED WORK

A good news headline is able to convey a concise summary of the news article [15], [16] and attract readers’ attention. The process of automatic headline generation has been a research focus in NLP. This section will present the previous work on NHG and the Malay language NHG.

2.1. News headline generation

In the early years, Gattani [17] categorized NHG into three broad approaches: ruled-based, statistics-based, and summarization-based. The rule-based approach detects and compresses important parts of an article using handcrafted linguistical rules such as hedge trimmer [16]. Even though this approach is simple yet lightweight, it is not able to discover the complex relationships in the text. Hence, the message in the headline deviates from the original content [18]. Meanwhile, the statistical-based approach uses statistical methods to discover the correlation between the words in the headlines and in the news content. Some of the early adoptions of this approach are naïve Bayes [10] and unsupervised topic discovery (UTD) by Dorr *et al.* [19]. The weakness of these models is the large training dataset requirement. On the other hand, the summarization-based approach regards headlines as very short summaries of content. Traditional text summarization approaches were adopted to produce a one-line sentence that can be regarded as a headline [20]. This approach requires multiple steps or combined steps of sentence selection and compression. The weakness of the approach is that the compression approach is not suitable for producing headlines that are only 10% or less than the original news articles. Besides that, the output may face contextual distortion due to the reused words.

Later, the NHG has been re-categorized into extractive and abstractive approaches [3]. Both the rule-based and statistical-based fall into the extractive approach. The abstractive approach, however, is able to maintain and focus on the main idea of the news when generating the headline [21]. The generated headline may not contain words or redundant words from the original news article, which is a common issue faced in the extractive approach. Sentence fusion or sentence compression techniques are some of the early adoptions of abstractive methods. The main issue faced by these approaches is the generated headlines were not able to convey the main idea of the news and were very lacking in grammatical structure.

The deep learning-based approach with the adoption of seq2seq has shown its success in text summarization [22]. This model consists of the encoder-decoder architecture [8], [9]. The encoder will compute the hidden state of each word sequence and the decoder will calculate the probability of each word in the vocabulary to generate the output sequence.

2.2. Low-resource language

LRL are also known as under-resourced languages. In NLP, these are referred as languages that have limited readily available linguistic-related materials and/or digital data. Usually, such languages face challenges, including limited lexical resources and sparse corpora. In this research, the focus is on the Malay language. The Malay language is a language under Nusantara in the Austronesia language family [23], [24]. Generally, this language has more than 290 million native speakers [24]. In the area of NLP, the research activities involving the Malay language are very limited. In the area of text summarization and NHG, most of the research mainly involves private datasets for experiments. The main sources of the news are from Berita Harian [25], [26], Utusan Malaysia [25], [26], and Bernama [25]–[27]. Generally, there are no readily public datasets available.

2.3. Recurrent neural network

As a class of artificial neural networks (ANN), recurrent neural networks (RNNs) are capable of processing sequential data by maintaining a hidden state that captures information from previous inputs. The nature of this design makes RNN suitable for sequential tasks in text generation [28]. Therefore, in seq2seq architectures, RNNs are used in both encoder and decoder to capture the sequential information from input and generate output in sequence.

The common RNN architectures include basic RNN, long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), and gated recurrent unit (GRU). The basic RNN is also called vanilla RNN. It works by storing and updating the hidden state at each time step. However, it has a weakness in handling long-term dependencies [29]. The LSTM is considered an advanced variant of vanilla RNN. It is designed to overcome the issue of long-term dependencies [30]. The LSTM cell has three gates, the input, output, and forget gates. These gates will manage the flow of the information from each time step. It can capture long dependencies and maintain information over longer sequences. The BiLSTM takes a step further by merging information from the past and the future in each time step [31]. This RNN is able to hold the information of the past and future using the forward state in a positive time direction and the backward state in a negative time direction. Research by Cho *et al.* [32], the GRU was proposed. This RNN is another variant of the vanilla RNN. It simplifies the gate design by using the reset and update gates, which will decide the amount of information to keep or discard from previous and current time steps.

2.4. Subword tokenization

In NLP-related tasks, word representation, also known as word embedding, plays a critical role. This process converts the input text into numerical vectors, which contain the semantic representation between the words and the context information of the text. The traditional word representation technique uses one-hot encoding which involves a 1 in the corresponding word position's index and 0 for the rest of the words' indices. This representation leads to sparse matrix formation in the embedding which is not computationally optimized [33]. Besides that, this representation is not capable of capturing the semantic relationship between words.

Meanwhile, the other classic types of embeddings adopt statistical probability approaches such as Word2Vec [33] and GloVe [34]. Both approaches are able to capture the word relationship and use the neighboring words to form meaningful word relationships. However, these two approaches are not capable of handling new words or unseen words from the input of the NLP tasks. Besides that, they need a large embedding for languages with complex morphological structures. To overcome the issues of rare words and OOV, subword embedding is introduced. This embedding will take an input word and split the word into smaller units such as character n-gram or linguistic subword units. Some of the notable works in morpheme segmentation are in [35]–[38].

3. PROPOSED METHOD

The proposed new MSAS method comprises of a hybrid of the extractive and abstractive summarization models. The MSAS is constructed with four major algorithmic elements, as shown in Figure 1. These four algorithmic elements are: i) leading sentence news TextRank; ii) Malay phoneme subword tokenizer, iii) Malay phoneme subword tokenizer subword embedding; and iv) seq2seq encoder and decoder. The MSAS model has adopted a hybrid design concept by binding the extractive and abstractive summarization approaches into one. The motivation for using this hybrid design is to overcome the common issue faced in the generic seq2seq training model when computing with long input sequences. As the training sequence length increases, the hardware memory requirement increases. Therefore, this hybrid approach aims to overcome the high memory requirement in the model training.

In the first computational element of MSAS, the extraction approach has been introduced to reduce the overall length of the input sequence to the training using leading sentence news TextRank, which helps to reduce the input by extracting salient news sentences from the input source [39]. Subsequently, the extracted salient news sentences' words into subword tokens are tokenized using the Malay phoneme subword tokenizer. Finally, the tokenized subwords are fed into the seq2seq model and trained with the Malay phoneme subword embedding.

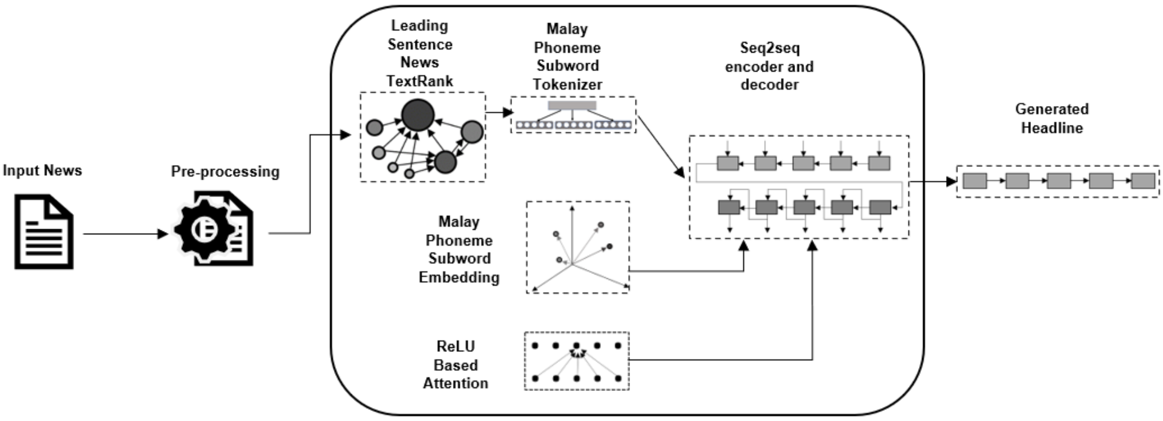
Input News
Pre-processing
Leading Sentence News TextRank
Malay Phoneme Subword Tokenizer
Seq2seq encoder and decoder
Generated Headline
Malay Phoneme Subword Embedding
ReLU Based Attention

Figure 1. MSAS

3.1. Leading sentence news TextRank

The TextRank algorithm is an unsupervised learning algorithm. In the beginning, this algorithm requires the input text to be transformed into a form of vector representation using the term frequency-inverse document frequency (TF-IDF) representation at the sentence pre-processing level. These vector representations will be fed into the TextRank algorithm to construct the relation between the nodes. In order to allow all the nodes with equal probability to be selected as the starting node at each iteration, the TextRank algorithm applies random probability to initialize all nodes at the beginning of each learning process. The relationship between the nodes is represented by the "strength" between the nodes. The strength is computed based on the words' similarities [39].

The leading sentence news TextRank adopted the TextRank approach that is customized for news article salient sentence selection. According to Tsann *et al.* [39] this algorithm will require setting 1 to the weight of node of the first sentence of the input text. The (1) is used to express the weight of a node.

$$TextRank(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} TextRank(V_j) \quad (1)$$

In the equation, the weight between the nodes is expressed as w_{ji} and the damping factor, d is set to 0.85.

3.2. Malay phoneme subword tokenizer

The Malay language is a type of agglutinative language. This language is full of morphemes. An example of a root word, "jangka" which means "expect" in English. The application of a prefix will become "menjangka" which means "to expect" in English. The application of a suffix will turn the word into "jangkaan" which means "expectations" in English. The application of the circumfix can turn the word into "dijangkakan" which means "expected" in English. The application of various phonemes has changed the root word into different meanings in the target translated language [40].

The application of traditional word embedding will require the representation of a single word that has multiple different applications of morphemes. This will increase the number of words in the embedding space. The application of stemming is not a practical approach as the context of a specific translated word may be different in meaning. Besides that, reducing the number of words in the word embedding space will also lead to OOV as some of the words will be removed during the word representation process.

To overcome the complex morphological structure in the Malay language compared to the English language, this study proposes a new Malay phoneme-based tokenization approach. This new tokenization approach will split a word into multiple subword tokens. Instead of algorithm-based tokenization, this approach will split a Malay word into subword tokens based on the four distinct syllable structures in the Malay language [41]. These syllables are based on the vowel, V, and consonant, C, then the four proposed syllable structures are subsequently: V, CV, VC, and CVC to be used to determine the subword tokens. Table 1 shows a set of samples of Malay words that are based on various syllabic structures as described above.

The proposed new tokenizer will perform the tokenization process to structure the input text into Malay phoneme subwords through an iterative process for each character in the input text until all the characters in each word have completed forming the common syllable structure. A special end-of-word symbol, '_,' will be added to the last subword token to indicate the end of a word. Thus, the new tokenization process is completed.

Table 1. Syllable structures of Malay words

Type of syllable	Syllable structure	Example word
Monosyllabic	CVC	*cik, sup, Jun*
	CV	*di, ke*
Dissyllabic	CV + CV	*peli, cari, lagu*
	V + CV	*abu, itu, ini*
	V + CVC	*ulat, ikat*
Trisyllabic	CV + CV + CV	*perasa, kerusi*
	CV + V + CV	*buaya, kuasa*
	CV + CV + V	*semua, ketua*
Tetrasyllabic	CV + CV + CV + CV	*daripada*
	CV + CVC + CV + VC	*perempuan*
	V + CV + CV + CV	*apabila*

3.3. Sequence-to-sequence

In 2014, Shwartz and Zhang [42] proposed the seq2seq model which consists of an encoder and a decoder. The encoder consists of a stack of RNN-based cells. The input sequence is fed into the encoder. Then, the input vector will be computed, and the internal state will be stored in the RNN-based cell. The (2) shows the computation of the hidden state h_i is computed using the weight function of the existing input x_t, previous state h_{t-1} and the network weight, W.

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

On the other side of the model, the decoder is formed using a stack of recurrent-based cells. The decoder will take in the final state of the encoder as initial input. Each time step t, the hidden state h_i is calculated using the output from the previous time step using (3).

$$h_t = f(W^{hh}h_{t-1}) \tag{3}$$

The output of each time step, y_t is computed using a Softmax function with the combination of the hidden state of the current time step and the respective weight W^S. The probability vector of the output y_t is shown in (4).

$$y_t = softmax(W^S h_t) \tag{4}$$

In the proposed seq2seq model, a single-layer RNN-based cell is used as an encoder, and a single-layer RNN-based cell as the decoder.

3.4. Attention mechanism

In 2015, Bahdanau *et al.* [7] first introduced the attention mechanism into the seq2seq model. The implementation will overcome the limitations of traditional seq2seq in handling long sequences. In the traditional seq2seq, the RNN-based cells must maintain all the hidden states for each time step of the input sequence until the end of the input sequence. Hence, this information retention leads to high memory and computational requirements. This becomes even more challenging in handling all the hidden states of long-range dependencies.

The attention mechanism will allow the model to only focus on the relevant parts of the input effectively without increasing the memory and computational requirements [43]. The decoder will selectively focus on the different parts of the input sequence during each decoding step based on the attention weight [44]. With the implementation of the attention mechanism, the seq2seq model will consist of three parts: the encoder, the decoder, and the attention layer. The attention layer contains three parts that include the alignment layer, attention weights, and context vector.

In the attention mechanism, the alignment layer plays an important role in computing the alignment scores between the input sequence and the output sequence of the seq2seq. These scores are also called attention scores or attention weights. The amount of these scores determines how much the focus is on a particular input sequence when generating a specific output sequence. The current state of the alignment score calculation will involve the hyperbolic tangent function that applies to the scores of the previous state h_{t-1} and previous state s_{p-1} as shown in (5).

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

In the proposed model, the original hyperbolic tangent function is replaced with a rectified linear unit (ReLU) function as shown in (6). This implementation aims to solve the vanishing gradient that commonly occurs in using the tanh function [45]–[47]. Besides that, it will allow more localized attention.

$$r_{rp} = v_a^T ReLU(W^{ss} s_{p-1} + W^{hh} h_{t-1}) \quad (6)$$

The attention weights for a target word and a set of source words are shown in (7).

$$\alpha_{tp} = \frac{\exp{(r_{rp})}}{\sum_{t=1}^{|x|} \exp{(r_{rp})}} \quad (7)$$

The attention weight α_{tp} and the hidden state h_t are used to compute the context vector c_p as shown in (8). These weights are obtained by multiplying the encoder states and the decoder hidden states.

$$c_p = \sum_{t=1}^{|x|} \alpha_{tp} h_t \quad (8)$$

The final distribution probability P_j is calculated by applying a softmax function as shown in (9).

$$P_j = softmax(W^s t_j) \quad (9)$$

4. EXPERIMENT AND RESULTS

Two RNN-based seq2seq models, the LSTM and BiLSTM seq2seq models are used as the experiment's baseline models to test against the new MSAS-tanh and MSAS-ReLU models. All these models were configured with the Bahdanau attention mechanism. This section details the dataset selection, text cleaning, input pre-processing, and the model configurations of the experiment.

4.1. Dataset

The dataset used in this paper is online Malay language news that are randomly collected from various news websites (i.e. Bernama.com, myMetro, and themalaysianinsight.com). These news articles were published between 2017 and 2019. About 98.9% of this news are hard news on current affairs. The remaining news are human interest soft news. The breakdown of the news genres is as mentioned: i) national news; ii) sports news; iii) economy and finance news; iv) world news; vi) entertainment and celebrity news; and vi) others such as education and lifestyles.

The average number of words in the news is 219.08 words and the longest article has 3,388 words. The average number of words in a headline is 3.83. The number of news used for training is 40,000 and 5,000 for testing. This train-test split of 90% and 10% will prevent the overfitting issue of the model training.

4.2. Text cleaning

The data pre-processing began with putting all the input text into lowercase. Then, the regular expression and the beautiful soup parser in Python were employed to remove any embedded HTML tags and Google Tags from the input news text to produce a clean input text representation. To further reduce the noise in the input text, punctuation marks and non-informative special characters were removed by using the regular expression. The data pre-processing was further enhanced with the removal of all non-textual-based characters using the regular expressions in Python.

4.3. Input pre-processing

Based on the experiment, the LSTM and BiLSTM seq2seq models used word-level tokenization to input the text into the embedding layer. Then, the post-padding was applied to the tokenized input of the models to ensure all the input had a length of 100 words. On the other hand, the MSAS-tanh and MSAS-ReLU models' input went through salient sentence extraction using the leading sentence news TextRank algorithm. Then, the extracted sentences went through the subword tokenization process using the Malay phoneme subword tokenizer to break each input word into subword tokens according to the tokenizer's syllable structures. Next, the post-padding was applied to the subword tokens to ensure all the input had a length of 100 words. Subsequently, all the inputs were fed into the embedding layer of the seq2seq model. The embedding layers for all the models were configured with 300 dimensions.

4.4. Model building

Based on the investigation of literature [7], [48], the LSTM-based seq2seq models are commonly used in text summarization. Hence, two seq2seq models: LSTM and BiLSTM have been selected as baseline models in the experiment. In the seq2seq model, an encoder and a decoder with an attention mechanism are employed. Each of the seq2seq models will contain a single RNN-based layer as encoder and a single layer of RNN-based as decoder. Both the encoder and decoder layers were configured with 300 hidden units per layer. To prevent overfitting issues, a dropout rate of 0.4 was set to the non-recurrent connection of the neurons during the training. At the same time, regularization is added in the layer to ignore some of the hidden state units using a 0.4 recurrent dropout rate. Besides that, these layers were configured with full sequence outputs and state returns within the architecture. These settings will improve the robustness and generalization ability of the models.

Subsequently, the Bahdanau *et al.* [7] attention mechanism was introduced to further improve the decoder's capability. It will allow the decoder to focus on the relevant parts of the input sequence when generating headlines. The attention mechanism will compute the attention score based on the encoder's hidden states according to the alignment to the decoder's current state. In the experiment, the attention mechanism's alignment score for LSTM, BiLSTM, and MSAS-tanh were computed using the hyperbolic tangent function. In the MSAS-ReLU model, the alignment score was computed using the ReLU function.

4.5. Training

To ensure the effectiveness of the model learning and prevent overfitting, all the models' training adopted the RMSprop optimizer with a default learning rate of 0.0001. The 'sparse_categorical_crossentropy' loss function was chosen as this function could handle the prediction of the next token problem. The minimize loss goal was implemented into the training process and the validation loss was monitored. The validation loss is the indicator of the model's generalization capability. To prevent extra training epochs and reduce overfitting risk, early stopping callback was introduced to the model training. All the model training was conducted over a maximum of 50 epochs with a batch size of 64 to balance the computational efficiency with memory usage. The experiments were performed using an Intel i7 PC with 32 GB of RAM, and an Nvidia GTX1080Ti GPU card with 12 GB of GPU RAM. All the models were developed using the Python tensorflow library running on the Ubuntu Desktop operating system.

4.6. Results and discussion

The objective of the following experiments is to compare the generated headlines of the proposed new MSAS models against the LSTM and BiLSTM seq2seq models. The experimental results were evaluated using the ROUGE evaluation metric, and the results demonstrated a significant improvement in the headline generation performance compared to other reported methods, BiLSTM and LSTM as delineated in Table 2. The proposed MSAS models have reported the following experimental results: i) the MSAS-tanh model achieved 16.62 ROUGE-1, 4.32 ROUGE-2, and 16.23 ROUGE-L, and ii) the MSAS-ReLU achieved 18.81 ROUGE-1, 5.07 ROUGE-2, and 18.35 ROUGE-L. According to the reported experimental results, both models consistently outperformed the standard LSTM-based models. Notably, both MSAS-based models recorded a minimum of 11.00 improvement in ROUGE-1 and ROUGE-L scores. Among all, the MSAS-ReLU exhibited the most significant improvement, as reported in Table 2 by its ROUGE scores.

Table 2. ROUGE scores for the models

	ROUGE-1			ROUGE-2				ROUGE-L		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Precision	Recall	F1
BiLSTM	5.63	4.73	4.98	0.82	0.69	0.71	0.82	5.57	4.69	4.93
LSTM	5.14	4.00	4.39	1.00	0.84	0.89	1.00	5.08	3.95	4.33
MSAS-tanh	18.73	16.02	16.62	5.01	4.12	4.32	5.01	18.29	15.65	16.23
MSAS-ReLU	21.29	18.04	18.81	5.88	4.96	5.07	5.88	20.79	17.60	18.35

Furthermore, Table 3 shows the average OOV rates for the BiLSTM, LSTM, MSAS-tanh, and MSAS-ReLU test models. Based on the test results, the MSAS-based models demonstrate significant improvement in headline generation capabilities compared to the other two models. According to Table 3, the MSAS-based models managed to reduce the OOV rates significantly. MSAS-based models achieved below 16% of the OOV rate with MSAS-tanh at 15.5% and RMAS-ReLU at 14.5%, respectively. These results show that MSAS-based models are capable of reducing the vocabulary limitation that often occurs in most of the seq2seq models that are widely used in headline generation. This notable improvement suggests that the proposed model can overcome the large vocabulary size requirement in the embedding layers of the model, which is helpful for the Malay language that has complex morphological structures. Besides that, the approach allows a more compact representation of the vocabulary.

Table 3. Model analysis for the models

Model	OOV rate (%)
BiLSTM	60.7
LSTM	74.7
MSAS-tanh	15.5
MSAS-ReLU	14.5

It is also reported in Table 2, when comparing the two MSAS models, the ReLU attention model demonstrates significant improvement in the numerical quantity of 2.19 for ROUGE-1, 0.75 for ROUGE-2, and 2.12 for ROUGE-L compared to MSAS-tanh. Additionally, the MSAS-ReLU model's OOV rate is also reduced by 1.00% and exhibits great computational efficiency that only requires 616.63 seconds per epoch for training as compared to the MSAM-tanh model, which requires an average of 624.6 seconds. These findings indicate the computational efficiency and effectiveness of the MSAS-ReLU model in inference and learning from data. Figure 2 shows the example of a summary produced by all the models in the experiment. From the sample, the output from the LSTM appears to have the OOV issue. The output of BiLSTM and MSAS-tanh were able to produce related headline but the context is not correct. The MSAS-ReLU model produced a more fluent headline.

I1: "*tanah merah mayat remaja lelaki yang terjun ke sungai jenob bukit bunga di sini rabu lalu ditemui dua kilometer dari lokasi mangsa dilaporkan hilang kelmarin mangsa yang berusia tahun ditemui anggota mencari dan menyelamat jam petang ketua polis daerah tanah merah deputi superintendan suzaimi mohamad berkata mayat mangsa ditemui tersangkut pada kayu di sungai berkenaan dan masih berpakaian lengkap …*" [49]
G: *mayat remaja tersangkut kayu*
BiLSTM: *mayat mayat dalam bilik tandas*
LSTM: update OOV OOV OOV
MSAS-tanh: *mayat remaja terjun sungai*
MSAS-ReLU: *mayat remaja ditemui*

I2: "*seramai ahli sindiket macau scam ditahan sejak awal tahun ini membabitkan kerugian rm juta setakat november selepas memperdaya mangsa di kedah ketua jabatan siasatan jenayah komersil kedah superintendan chan teck paing berkata angka itu juga menunjukkan peningkatan berbanding tahun lalu …*" [50]
G: *ramai orang kedah kena tipu macau scam*
BiLSTM: *polis cari lelaki warga asing ditahan*
LSTM: OOV OOV OOV
MSAS-tanh: *sindiket macau scam ditahan*
MSAS-ReLU: *sindiket macau scam dicekup*

Figure 2. Example headline produced. I1 and I2 are the input, and G is the actual headline

Based on the findings, the MSAS-ReLU model demonstrated superior performance, which can be attributed to several factors. Adopting the ReLU attention score mechanism likely helps alleviate the vanishing gradient problem and maintains better gradient flow in the backpropagation in the model. Besides that, the ReLU function is more computationally efficient compared to the Tanh function due to its simpler mathematical operations that directly translate into the same expectation in computation. Overall, we see a reduction in OOV rates, which can be related to the implementation of the phoneme-based custom subword embedding. This approach effectively addresses the vocabulary limitation through the decomposition of words into manageable subwords.

5. CONCLUSION

This paper presents a comparison between contemporary RNN-based models and the newly proposed MSAS models to address the generation of Malay language news headlines. Due to the limitation of the existence of an open Malay language news dataset available for the experiment, the two RNN-based models were set to the same property dataset that was collected specifically for this research. Based on the empirical evaluation of the experiment result, the newly proposed model, MSAS-ReLU, demonstrated superior performance over the BiLSTM, LSTM, and MSAS-tanh models in the experiments. The MSAS-ReLU model shows its algorithmic learning and inference effectiveness, particularly in the areas of improvement in ROUGE scores and reduction of the OOV rate. These notable experimental achievements clearly highlight the proposed new MSAS hybrid extractive and abstractive summarization models implemented together with phoneme-based subword embeddings are able to generate more accurate news headlines even though the proposed MSAS models are learning and inference in datasets with unbalanced constraints of vocabulary limitation. In short, the proposed new MSAS models have significantly solved the complex morphological structure and low-resource training issue commonly found in Malay language-based computational linguistic applications. The MSAS model's experimental success clearly indicates that it is a guideline for other researchers to implement future Malay language NHG algorithms using MSAS as the foundation model. Also, the achievement of the MSAS role model solidifies a new benchmark for Malay MHG tasks. This research paves the way for future Malay language NHG exploration to invent other advanced and modern seq2seq models, specifically in the linguistic application using the Malay language.

ACKNOWLEDGEMENTS

This research was fully funded by the Yayasan UTP-PRG Cycle 2022 (015PBC-005) and Department of Computer and Information Science at Universiti Teknologi PETRONAS. The authors would like to thank Yayasan UTP-PRG for supporting this work.

REFERENCES

[1] G. M. Montejo and T. Q. Adriano, "A critical discourse analysis of headlines in online news portals," *Journal of Advances in Humanities and Social Sciences*, vol. 4, no. 2, Apr. 2018, doi: 10.20474/jahss-4.2.2.

[2] K. Kaikhah, "Automatic text summarization with neural networks," in *2004 2nd International IEEE Conference on 'Intelligent Systems'*, IEEE, 2004, pp. 40–44. doi: 10.1109/IS.2004.1344634.

[3] O. Tilk and T. Alumäe, "Low-resource neural headline generation," *arXiv-Computer Science*, pp. 1-7, Jul. 2017, doi: 10.48550/arXiv.1707.09769.

[4] R. Zhang, J. Guo, Y. Fan, Y. Lan, and X. Cheng, "Structure learning for headline generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 9555–9562, Apr. 2020, doi: 10.1609/aaai.v34i05.6501.

[5] B. K. Boguraev and M. S. Neff, "Discourse segmentation in aid of document summarization," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, IEEE Comput. Soc, 2000, doi: 10.1109/HICSS.2000.926687.

[6] Z. Li, J. Wu, J. Miao, and X. Yu, "News headline generation based on improved decoder from transformer," *Scientific Reports*, vol. 12, no. 1, Jul. 2022, doi: 10.1038/s41598-022-15817-z.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv-Computer Science*, pp. 1-15, Sep. 2014, doi: 10.48550/arXiv.1409.0473.

[8] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003, doi: 10.1109/TNN.2003.820440.

[9] R. Nallapati, B. Zhou, C. D. Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *The 20th SIGNLL Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 280–290, doi: 10.18653/v1/K16-1028.

[10] M. Banko, V. O. Mittal, and M. J. Witbrock, "Headline generation based on statistical translation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics-ACL '00*, Morristown, NJ, USA: Association for Computational Linguistics, 2000, pp. 318–325, doi: 10.3115/1075218.1075259.

[11] D. Zajic, B. Dorr, and R. Schwartz, "Automatic headline generation for newspaper stories," in *Proceedings of the ACL-2002 Workshop on Text Summarization*, 2002.

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

[13] U. Kumar, V. Singh, C. Andrew, S. Reddy, and A. Das, "Consonant-vowel sequences as subword units for code-mixed languages," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018, pp. 8103–8104, doi: 10.1609/aaai.v32i1.12193.

[14] O. Kwon, D. Kim, S. R. Lee, J. Choi, and S. K. Lee, "Handling out-of-vocabulary problem in hangeul word embeddings," in *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2021, pp. 3213–3221, doi: 10.18653/v1/2021.eacl-main.280.

[15] Ayana *et al.*, "Recent advances on neural headline generation," *Journal of Computer Science and Technology*, vol. 32, no. 4, pp. 768–784, Jul. 2017, doi: 10.1007/s11390-017-1758-3.

[16] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer," in *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 1–8, doi: 10.3115/1119467.1119468.

[17] A. K. Gattani, "Automated natural language headline generation using discriminative machine learning models," *M.Sc. Thesis*, School of Computing Science, Simon Fraser University, Burnaby, Canada, 2017.

[18] M. Hassel and N. Mazdak, "FarsiSum," in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages-Semitic '04*, Morristown, NJ, USA: Association for Computational Linguistics, 2004, doi: 10.3115/1621804.1621826.

[19] B. Dorr, D. Zajic, and R. Schwartz, "Cross-language headline generation for Hindi," *ACM Transactions on Asian Language Information Processing*, vol. 2, no. 3, pp. 270–289, Sep. 2003, doi: 10.1145/979872.979878.

[20] C. A. Colmenares, M. Litvak, A. Mantrach, and F. Silvestri, "HEADS: headline generation as sequence prediction using an abstract feature-rich space," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 133–142, doi: 10.3115/v1/N15-1014.

[21] I. Mani, *Automatic Summarization*, Amsterdam: John Benjamins Publishing Company, 2001, doi: 10.1075/nlp.3.

[22] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 93–98, doi: 10.18653/v1/N16-1012.

[23] N. S. Karim, F. M. Onn, H. H. Musa, and A. H. Mahmood, *Hall grammar third edition (*in Malay*: Tatabahasa dewan edisi ketiga)*, Kuala Lumpur, Malaysia: Dewan Bahasa Dan Pustaka, 2015.

[24] N. M. Awal, K. A. Bakar, N. Z. A. Hamid, and N. H. Jalaluddin, "Morphological differences between bahasa melayu and english: constraints in students' understanding," *School of Language Studies & Linguistics, Universiti Utara Malaysia*, pp. 1-11, 2007.

[25] S. A. M. Noah, N. M. Ali, and M. S. Hasan, "Determining features of news headline in Malay news document (in Malay: *Penentuan fitur bagi pengekstrakan tajuk berita akhbar bahasa melayu*)," *Journal of Language Studies*, vol. 18, no. 2, pp. 154–167, 2018, doi: 10.17576/gema-2018-1802-11.

[26] S. Alias, S. K. Mohammad, G. K. Hoon, and M. S. Sainin, "Understanding human sentence compression pattern for Malay text summarizer," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, IEEE, Mar. 2018, pp. 1–6, doi: 10.1109/INFRKM.2018.8464788.

[27] M. S. Hasan, S. A. M. Noah, and N. M. Ali, "Malay text features for automatic news headline generation," *Journal of Theoretical and Applied Information Technology*, vol. 76, no. 1, pp. 36–41, 2015.

[28] R. Siddalingappa and K. Sekar, "Bi-directional long short term memory using recurrent neural network for biological entity recognition," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 89–101, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp89-101.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[30] E. Mathew and S. Abdulla, "The LSTM technique for demand forecasting of e-procurement in the hospitality industry in the UAE," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 757–765, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp757-765.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[32] 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)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[33] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in Neural Information Processing Systems*, 2001.

[34] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[35] S. Nießen and H. Ney, "Improving SMT quality with morpho-syntactic analysis," in *Proceedings of the 18th conference on Computational linguistics*, Morristown, NJ, USA: Association for Computational Linguistics, 2000, doi: 10.3115/992730.992809.

[36] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-EACL '03*, Morristown, NJ, USA: Association for Computational Linguistics, 2003, doi: 10.3115/1067807.1067833.

[37] S. Virpioja, J. J. Väyrynen, M. Creutz, and M. Sadeniemi, "Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner," in *Proceedings of the Machine Translation Summit XI*, pp. 491–498, 2007.

[38] D. Stallard, J. Devlin, M. Kayser, Y. K. Lee, and R. Barzilay, "Unsupervised morphology rivals supervised morphology for Arabic MT," in *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012-Proceedings of the Conference*, pp. 322–327, 2012.

[39] P. Y. Tsann, Y. K. Hooi, M. F. Hassan, and M. T. Y. Wooi, "Leading sentence news TextRank," in *2021 International Conference on Intelligent Cybernetics Technology and Applications (ICICyTA)*, IEEE, Dec. 2021, pp. 92–95, doi: 10.1109/ICICyTA53712.2021.9689186.

[40] L. W. Lee, H. M. Low, and A. R. Mohamed, "A comparative analysis of word structures in Malay and English storybooks social sciences and humanities a comparative analysis of word structures in Malay and English children's stories," *Pertanika Journal of Social Science*, pp. 67–84, 2013.

[41] L. W. Lee, "Development and validation of a reading-related assessment battery in Malay for the purpose of dyslexia assessment," *Annals of Dyslexia*, vol. 58, no. 1, pp. 37–57, Jun. 2008, doi: 10.1007/s11881-007-0011-0.

[42] S. S. -Shwartz and T. Zhang, "Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization," *Mathematical Programming*, vol. 155, pp. 105–145, Jan. 2016, doi: 10.1007/s10107-014-0839-0.

[43] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, Apr. 2018, doi: 10.1016/j.neucom.2018.01.007.

[45] 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.

[46] T. Szandała, "Review and comparison of commonly used activation functions for deep neural networks," in *Studies in Computational Intelligence*, 2021, pp. 203–224, doi: 10.1007/978-981-15-5495-7_11.

[47] Y. T. Phua, S. Navaratnam, C.-M. Kang, and W.-S. Che, "Sequence-to-sequence neural machine translation for English-Malay," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 658–665, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp658-665.

[48] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, “Neural abstractive text summarization with sequence-to-sequence models,” *arXiv-Computer Science*, pp. 1-35, Dec. 2018, doi: 10.48550/arXiv.1812.02303.

[49] H. Digital, “The body of a teenager stuck in a tree (in Malay: *Mayat remaja tersangkut kayu*),” *HM Digital*. Accessed: May 16, 2024. [Online]. Available: https://www.hmetro.com.my/amp/mutakhir/2017/11/282908/mayat-remaja-tersangkut-kayu

[50] Z. Zulkiffli, “Many Kedah people have been deceived by the macau scam (in Malay: *ramai orang kedah kena tipu macau scam*),” *HM Digital*, Kuala Lumpur. Accessed: May 16, 2024. [Online]. Available: https://www.hmetro.com.my/mutakhir/2017/12/295037/ramai-orang-kedah-kena-tipu-macau-scam

BIOGRAPHIES OF AUTHORS

Yeong Tsann Phua an alumnus of Universiti Teknologi Malaysia (UTM) with a B.Sc. in Computer with Education (1997) and Universiti Malaya (UM) with a Master in Computer Science (2006), is currently a Ph.D. student at Universiti Teknologi PETRONAS (UTP). His research focuses on machine learning, deep learning, natural language processing, and computer vision, with a current emphasis on text summarization in the Malay language. He can be contacted at email: yeong_17008256@utp.edu.my.

Kwang Hooi Yew received the Ph.D. degree from Universiti Teknologi PETRONAS, Malaysia. He is currently a Senior Lecturer with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS. His research interests include the semantic web, knowledge representation, and formal language. He can be contacted at email: yewkwanghooi@utp.edu.my.

Mohd Fadzil Hassan holds a Bachelor of Engineering (B.Eng.) in Electrical and Electronics Engineering, He received the B.Sc. degree (cum laude) in Computer Information Systems from Colorado State University, USA, in 1999, and the M.Sc. degree in artificial intelligence and the Ph.D. degree in informatics from the University of Edinburgh, U.K., in 2001 and 2007, respectively. He was the former Dean of the Centre for Graduate Studies, Universiti Teknologi Petronas (UTP), where he is currently the Director of the Institute of Autonomous Systems. He is also an alumnus of the Malay College Kuala Kangsar (MCKK). He is senior member at IEEE. He has been involved in authoring more than 100 indexed publications. His research interests include artificial intelligence, multiagent systems, and service-oriented architecture (SOA). He is also actively involved in international collaborative research, particularly with universities from the Middle East, South Korea, and the ASEAN region. He was an Executive Committee Member of the IEEE Computer Society Malaysia, in 2018. He can be contacted at email: mfadzil_hassan@utp.edu.my.

Matthew Teow Yok Wooi graduated with a B.Sc. in Electronic and Electrical Engineering from Robert Gordon University, UK, an M.Eng. in Electrical Engineering from Universiti Teknologi Malaysia, Malaysia, and a Ph.D. in Engineering from Multimedia University, Malaysia. He is a lecturer at University Partnership (Coventry University), PSB Academy, Singapore. He has published more than 30 papers in conferences and journals. He has extensive interests in artificial intelligence and scientific computing applied to computational learning and inference methods. He can be contacted at email: matthew.teow@psb-academy.edu.sg.