JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
http://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Alias, S., Sainin, M. S., & Mohammad, S. K. (2021). A syntactic-based sentence validation technique for Malay text summarizer. *Journal of Information and Communication Technology, 20*(3), 329-352. https://doi.org/10.32890/jict2021.20.3.3

A Syntactic-based Sentence Validation Technique for Malay Text Summarizer

[1]Suraya Alias, [2]Mohd Shamrie Sainin,
[3]Siti Khaotijah Mohammad
[1&2]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia
[3]School of Computer Sciences, Universiti Sains Malaysia, Malaysia

suealias, shamrie@ums.edu.my
sitijah@usm.my

Received: 19/10/2020 Revised: 10/1/2021 Accepted: 31/5/2021 Published: 11/6/2021

ABSTRACT

In the automatic text summarization domain, a sentence compression technique is applied to the summary sentence to remove unnecessary words or phrases. The purpose of sentence compression is to preserve important information in a sentence and to remove unnecessary ones without sacrificing the sentence's grammar. The existing development of Malay natural language processing tools is still under study with limited open access. The issue is the lack of a benchmark dataset in the Malay language to evaluate the quality of the summaries and to validate the compressed sentence produced by the summarizer model. Therefore, this paper outlines a syntactic-based sentence validation technique for Malay sentences by referring to the Malay grammar pattern. In this work, a new derivation set of

syntactic rules based on the Malay main word class was proposed to validate Malay sentences that underwent the sentence compression procedure. This paper used the Malay dataset of 100 new articles covering the natural disaster and events domain to find the optimal compression rate and its effect on the summary content. An automatic evaluation using the benchmark ROUGE toolkit produced a result with an average F-measure of 0.5826 and an average recall value of 0.5925 with an optimum compression rate of 0.5 confidence *conf* value. Furthermore, a manual summary evaluation by a group of Malay experts on the grammaticality of the compressed summary sentence produced a good result of 4.11 and a readability score of 4.12 out of 5. This depicts the reliability of the proposed technique to validate Malay sentences with promising summary content and readability results.

Keywords: Malay text summarization, Sentence compression, Syntactic rules, POS, Parser.

INTRODUCTION

In the domain of automatic text summarization, sentence compression can be referred to as a sentence-level summarization procedure. Sentence compression aims to delete an insignificant phrase from a summary sentence and preserves the significant ones by validating the grammaticality of the new compressed sentence. The application of sentence compression can help to improve the consistency and readability of the summary.

Related studies in automatic text summarization, sentence compression, and existing benchmark datasets are mostly in English. The main problem is the lack of a benchmark dataset in the Malay language to evaluate the summaries produced automatically by the summarizer model (Jusoh et al., 2011; Zamin & Ghani, 2010). This issue has prompted the development of a Malay summary corpus to understand human compression patterns and selected features in producing a summary. These findings were then utilized in Alias (2018) to develop a Malay text summarizer model named MYTextSumCOMP with Pattern-Growth sentence compression technique.

It is a challenge to apply a syntactic-based technique on Malay sentence validation due to limited tools for natural language processing (NLP) such as Malay Part-of-Speech (POS) tagger and parser. Furthermore, such tools are not available for public access. Several previous research in Malay POS tagger used the formal language corpus of news and non-news data (Alfred et al., 2013; Mohamed et al., 2011; Xian et al., 2016). In contrast, recent works by Ariffin and Tiun (2018) have shifted to social media texts, where the authors highlighted the issue of informal language used in Tweet texts and the need for a large training of Malay corpus to be set up. They adopted a probabilistic supervised approach named QTAG with accuracy results of 90 percent.

Recently, Hamza et al. (2019) experimented using the thematic role approach to discover the relationship of each word and its role to preserve the meaning of a sentence. Their focus was on the short essay assessment dataset, and they highlighted that the subject-predicate rule had the limitation of not considering the sentence usage context. However, as sentence validation is performed on a summary sentence, preservation of the sentence meaning has been taken into consideration in the present study's text summarization task.

This study focuses on evaluating the performance of the proposed syntactic-based Malay sentence validation technique. It is implemented on the existing MYTextSumCOMP model by Alias (2018) – a model for Malay text summarization based on Pattern-Growth technique. To validate a Malay sentence, four basic grammar patterns (*pola ayat dasar*) are used as a reference, based on the work of Ab Aziz et al. (2006). This paper's approach consists of two phases, which are the shallow POS tagging by word class, with and phrase structure and parsing. The technique is evaluated on the summary produced by the MYTextSumCOMP model using sentence validation analysis, automatic evaluation using ROUGE by Lin (2004), and manual grammatical evaluation by a panel of experts.

RELATED WORKS

To date, research and development related to Malay NLP are still on-going and struggling to be set as a benchmark. This section highlights existing Malay works such as POS tagger and parser based on syntactic rules.

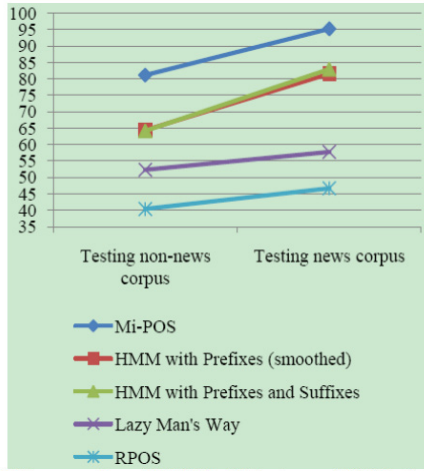
Malay Natural Language Processing

Related works in Malay POS tagger have been around in the literature, where techniques such as the statistical approach by Mohamed et al. (2011), rule-based by Alfred et al. (2013), syntactic data-driven by Mohd Don (2010), and Malay-English translation POS by Zamin et al. (2012) have been proposed. In Mohamed et al. (2011), their statistical approach used a Trigram Hidden Markov model (HMM) technique that utilized the sentence context information but relied on large training data and computation. On the other hand, a rule-based POS approach (RPOS) by Alfred et al. (2013) produced a good performance in predicting main POS word categories. Nevertheless, the issue arises where it requires many handcrafted rules that affect the accuracy of a non-main POS category.

MALEX by Mohd Don (2010) is a syntactic data-drive POS model that referred to syntactic trees wherein much effort was needed with comprehensive training data. In contrast, a Lazy Man's approach by Zamin et al. (2012) took advantage of translating from Malay to English in POS tagging. However, their approach resulted in low accuracy due to incorrect mapping in language structure. The current work by Xian et al. (2016) performed a benchmarking experiment on previous Malay POS approaches. Their Mi-POS used machine learning probabilistic methods for the POS tagging by referring to the training corpora. The Mi-POS results depicted an accuracy value of 95.16 percent using the same word corpora. However, it resulted in a decreased performance of 81.12 percent when tested using different ones. Figure 1 illustrates the accuracy results of the Mi-POS technique against the existing methods (HMM, Lazy Man's Way, and RPOS) as mentioned earlier using two different testing corpora of news and non-news.

Figure 1

The Accuracy Results of the Mi-POS Technique Against the Existing Methods Using Two Different Testing Malay Corpora (news and non-news) (Xian et al., 2016).



In contrast with Malay news and non-news corpora, a recent work in Malay POS tagger by Ariffin and Tiun (2018) experimented using social media texts by adopting a probabilistic supervised approach named QTAG. They reported an average accuracy result of 90 percent for a normalized dataset and 88.8 percent for an unnormalized dataset. Nevertheless, to achieve higher accuracy, it was noted that a larger training corpus in informal Malay language with different dialects (i.e., Negeri Sembilan, Kelantan and Perlis) is in dire need of future improvement.

Syntactic Based

Most linguistic rules that refer to syntactic and discourse knowledge of a sentence have been the basis for sentence compression technique grammar validation. It refers to heuristic knowledge extracted from humans where the training corpus is developed to understand better how humans construct and remove certain phrases. This approach has been the reference of the present study, whereby prior and current

works in English were adopted (Jing, 2000; Wang et al., 2013; Zajic et al., 2007). Jing (2000) developed an Automatic Sentence Reduction System that referred to rules from a syntactic parse tree. A phrase will be removed if: 1) the word/phrase is not required grammatically; 2) it is a non-significant phrase based on word scoring; and 3) the probability of removal using Naïve Bayes classifier from human heuristic knowledge is high. Nevertheless, this paper will highlight the expensive resource dependency, where the approach refers to the lexicon and WordNet database, which is also a challenge in the Malay language.

Following that, Zajic et al. (2007) implemented an iterative compression procedure of "parse-and-trim" to each constituent in a sentence. They also checked for the valid subject and predicate phrases so that the compressed sentence conformed to the sentence linguistic rules. The current work in multi-document summarization by Wang et al. (2013) referred to syntactic rules. Based on the syntactic parse tree, they developed a compression model to learn the most optimum compression by using a beam search decoder. This approach is similar to the current study's optimized confidence *conf* value of Frequent Eliminated Pattern (FASPe). In this paper, FASPe is a term that is frequently being eliminated by human experts when compressing a sentence derived from human compression patterns.

One of the common issues highlighted by both Malay and English syntactic-based approaches is the ambiguity in the structure of sentences produced by the parse tree. To overcome this, Hiloh et al. (2018) proposed a bottom-up parsing technique using the Cocke–Younger–Kasami (CYK) algorithm. The statistical parser calculated the probability of the most suitable parse tree by referring to the vocabulary and Malay grammar rules in the corpus. Their evaluation results stated an average of 97.1 percent for the correct proposed parse tree that was experimented by using simple Malay language sentences. Their findings reflected that the main reasons for incorrect sentence parsing were due to multiple POS tagging for certain words and the sentence structure that could be represented using more than one pattern, which is also in line with this research's findings.

Hamza et al. (2019) highlighted that the subject-predicate rule implementation in the Malay language had the limitation of not

considering the sentence usage context. Therefore, they proposed a thematic role approach to discover the relationship of each word and its role to preserve sentence meaning. They labeled each Malay word according to their thematic roles such as agent and patient, following the subject and predicate-argument. Their focus was on the short essay assessment dataset, whereas the current study is focused on the compressed Malay summary sentence through which the preservation of the sentence meaning has been taken into consideration and evaluated in the text summarization task.

A SYNTACTIC-BASED SENTENCE VALIDATION TECHNIQUE

The MyTextSumCOMP model represents important terms in a sentence using the Frequent Adjacent Sequential Pattern (FASP) weightage and adopts Greedy strategy for the sentence selection procedure to automatically produce a summary (Alias, 2018). The performance of MyTextSumCOMP model yielded a significant precision result of 0.5925 when tested against the Baseline summarizer model using the Malay dataset of 100 articles. The sentence validation procedure in the MyTextSumCOMP model inspected the compressed sentences to see if they conformed to Malay grammar patterns before they could be selected as a summary sentence. The proposed sentence compression technique was inspired by the Sequential Pattern Mining Pattern-Growth technique that consisted of three main steps: 1) Sentence segmentation; 2) Conquer significant segment(s); and 3) Syntactic-based sentence validation.

The essence of the proposed sentence compression technique implemented the divide-and-conquer strategy by removing unimportant segments from a sentence. The compression was performed based on the confidence *conf* value, in which the conditional probability ranged from 0.0 to 1.0 from the discovered Frequent Eliminated Pattern (FASPe). A FASPe with a high *conf* value depicted a set of terms that were frequently being eliminated by human experts when compressing a sentence. In contrast, the low ones indicated less eliminated frequent terms. During the sentence compression, the important terms were

preserved using the FASP representation and the sentence scoring features found in each sentence.

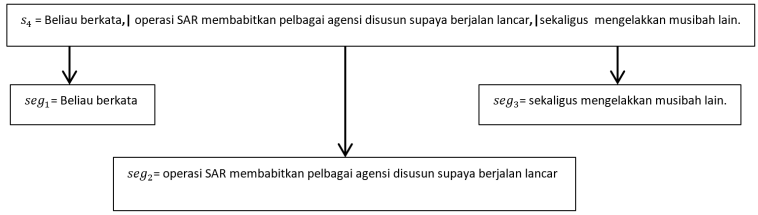
Sentence Segmentation

First, the sentences were divided into smaller segments to mark which words or phrases were eligible to be removed. Based on studies by Conroy et al. (2006) and Tran et al. (2015), delimiters, such as the commas, periods, and the beginning of a sentence, can be used to identify words or phrases to be removed in a sentence. As stated in Nik Safiah Karim et al. (2008), the usual practice for a comma (,) to be present at the beginning of a sentence indicated the presence of a discourse marker. In the proposed technique, for each article to be summarized, each source sentence was divided based on the comma (,) delimiter.

Given a source sentence *S* "*Beliau berkata, operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancar, sekali gus mengelakkan musibah lain*", the sentence was divided into three segments after the sentence segmentation procedure, as depicted in Figure 2.

Figure 2

The Sentence Segmentation Procedure for Malay Sentence Compression



Conquer Significant Segment(s)

After the source sentence was divided into segments, the technique proceeded to conquer only significant segments by referring to the frequent pattern weightage FASP that was found in each segment of the sentence. In contrast, the FASPe weight indicated the frequently

eliminated terms that matched in each sentence segment. The removal indicator was used to check if FASPe was greater than FASP to decide on the removal process of the respective segment. Table 1 illustrates the phrase removal process for a Malay sentence in the MyTextSumCOMP model.

Table 1

The Phrase Removal Process for a Malay Sentence in MyTextSumCOMP Model.

seg_i	Sentences	seg_i length	Matched FASPe	Matched FASP	FASPe Weight	FASP Weight	Removal Indicator
seg_1	***Beliau berkata***	2	*beliau, berkata*	*berkata*	1.0	0.5	**1**
seg_2	*operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancar*	9		*sar, agensi*	0.0	0.2	**0**
seg_3	***sekali gus mengelakkan musibah lain***	5	*lain*		0.2	0.0	**1**

Syntactic-based Sentence Validation

After the sentence compression procedure, the newly compressed sentence was validated to avoid any grammatical error. The sentence validation technique consisted of two phases: 1) Shallow POS tagging and 2) Phrase structure and parsing.

1) Shallow POS Tagging

First, a shallow POS tagging procedure was performed using the four main POS word classes in Malay, i.e. *Kata Nama* (KN), *Kata Kerja* (KK), *Kata Adjektif* (KA) and *Kata Tugas* (KT). In this work, the Malay POS sources were derived from Kamus Multimedia Bahasa Melayu - Bahasa Inggeris (KSMI) that consisted of 18,395 Malay tagged words. Figure 3 depicts the shallow Malay POS tagging procedure for the newly compressed sentence.

Figure 3

Shallow POS Tagging Procedure for the Newly Compressed Malay Sentence.

operasi/**KN** SAR/**KN** membabitkan/**KK** pelbagai/**KA** agensi/**KN** disusun/**KK** supaya/**KT** berjalan/**KK** lancar/**KA** ./.

2) Phrase Structure and Parsing

The subject and predicate phrases construct a basic sentence syntax. A subject is usually a noun phrase or a *Kata Nama* (KN) word, or other related noun elements. Meanwhile, a predicate describes the action of the subject using a group of other phrases. To validate a Malay sentence, four basic grammar patterns (*pola ayat dasar*) is used as a reference (Ab Aziz et al., 2006; Nik Safiah Karim et al., 2008; Omar 1998). Table 2 depicts the four basic Malay language grammar patterns with some examples. The subject phrase is formed by the *Frasa Nama* (FN), while the predicate phrase can be one of the FN, *Frasa Kerja* (FK), *Frasa Adjektif* (FA), or *Frasa Sendi* (FS) class.

Table 2

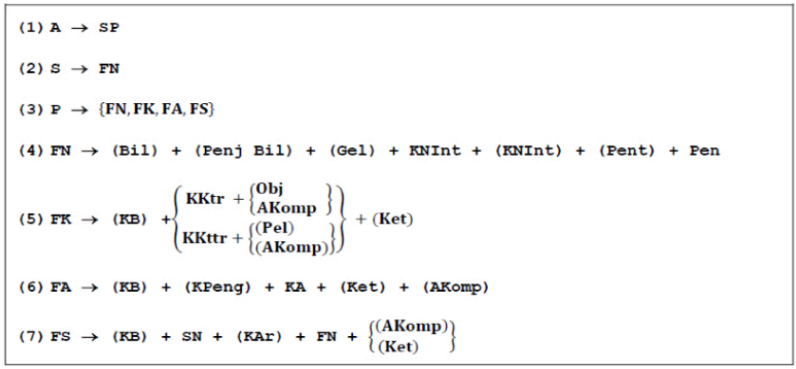
Malay Basic Grammar Pattern (Pola Ayat).

seg_i	Sentences	seg_i length	Matched FASPe	Matched FASP	FASPe Weight	FASP Weight	Removal Indicator
seg_1	***Beliau berkata***	2	*beliau, berkata*	*berkata*	1.0	0.5	**1**
seg_2	*operasi SAR membabitkan pelbagai agensi disusun supaya berjalan lancar*	9		*sar, agensi*	0.0	0.2	**0**
seg_3	***sekali gus mengelakkan musibah lain***	5	*lain*		0.2	0.0	**1**

The main list of POS elements used in the construction of a Malay phrase referred to in this study is shown in Figure 4.

Figure 4

Malay Language POS Elements by Nik Safiah Karim et al. (2008).



To create the syntactic-based validation rules, the POS elements in Figure 4 were analyzed and categorized into their main word class (KN, KT, KK, and KSN). The new syntactic Malay phrase derivation set is given in Table 3. For instance, for an FN, the core POS elements should have the KN class (Gel, KNInt) alongside the KT class, which included POS elements, i.e., Penj Bil and Pen. Therefore, a new set of derivation for an FN in the subject phrase can be written as FN → {KN, KT}.

Another example is for an FA containing KA class and KT class POS elements (e.g., KB, KPeng, and Ket). This study suggested a predicate phrase; an FA could be constructed by a set of POS sequences of the KA and KT word class. Therefore, a new set of derivation for a FA can be written as FA → {KA, KT}.

Table 3

A new Syntactic Malay Phrase Derivation Set Based on Malay Word Class

Malay phrase and POS elements word class	New syntactic Malay phrase derivation set
FN → (Bil)/KT + (Penj Bil)/KT + (Gel)/KN + KNInt/KN + (KNInt)/KN + (Pent)/KT + Pen/KT	∴ FN → {KN, KT}
FK → (KB)/KT + { KKtr/KK + { Obj/KN, AKomp/KT } , KKttr/KK + { (Pel)/KT, (AKomp)/KT } } + (Ket)/KT	∴ FK → {KK, KT, KN(Object)}
FA → (KB)/KT + (KPeng)/KT + KA/KA + (Ket)/KT + (AKomp)/KT	∴ FA → {KA, KT}
FS → (KB)/KT + SN/KSN/KT + (KAr)/KT + FN/KT, KN + { (AKomp/KT), (Ket)/KT }	∴ FS → {KSN, KT, KN}

The proposed sentence validation algorithm is described in Algorithm 1 (refer to Appendix), consisting of three modules to validate the subject phrase, predicate phrase, and the Malay sentence's syntax. First, the *msValidation(ms)* algorithm identified and validated the subject phrase. The sentence was then divided into segments of POS elements using KN as delimiters. Based on the new derivation, a valid subject phrase should consist of an FN with a set of sequences from KN and KT word classes (refer to Table 3). Normally, a Malay sentence will not start with either KK or KA in the subject phrase (unless it is a KN). Therefore, if the KK or KA word class was found in the subject phrase of the sentence, the sentence was considered invalid.

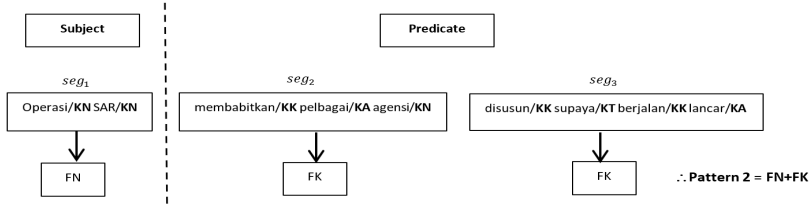
Later, in the next segment, the algorithm validated the predicate phrase by checking the new derivation set of the FN, FK, FA, or FS phrases. Recursively, the algorithm searched the core word class elements used to construct the phrase based on the new Malay phrase derivation set. For instance, to classify a valid FK phrase, the segment was checked for POS elements that were tagged under the KK word class. The final sentence validation module was to examine whether the syntax of the Malay sentence *ms* had valid

subjects and predicates before classifying it to one of the Malay grammar pattern groups.

If the Malay sentence *ms* did not conform to these rules, the current study would revert and optimize the sentence compression process again. Figure 5 illustrates the shallow POS tagging procedure for a Malay sentence by word classes and the output from the algorithm that classified the Malay grammar pattern (*pola ayat*). It was found that the newly compressed sentence was valid and majorly conformed to FN + FK, the Malay grammar pattern 2 (*Pola 2*). The final grammar evaluation for the Malay sentences was performed during the manual summary evaluation.

Figure 5

The Shallow POS Tagging Procedure for Malay Sentence by Word Class and the Malay Grammar Pattern Classification



RESULTS AND DISCUSSION

In this study, the performance of the sentence validation technique was based on the sentence validation analysis, automatic evaluation using ROUGE, and manual grammatical evaluation from a panel of experts. The chosen panels comprised secondary school Bahasa Melayu teachers who were native speakers with 15 years and above of teaching experience.

This research employed a new Malay dataset of 100 articles covering the natural disaster and events domain to find the optimal compression value and its effect on the summary content. The average word per article was 407, with a total count of 40,917. From the total of 1,883 sentences, 533 sentences were compressed by the system.

Sentence Validation Analysis

Table 4 depicts the number of invalid sentences identified from the total number of compressed sentences by each FASPe *conf* value. The findings showed that using a low *conf* value, i.e., 0.2, affected the grammar quality of the compressed sentences. Out of 551 compressed sentences, 54 were found invalid by the proposed sentence validation technique. From the analysis, FASPe with a low *conf* contained a set of terms that belonged to the sub-type kth in the KT word class, such as "*yang*" and "*ialah*". Therefore, removing the segments with these sets of terms vigorously affected the grammaticality of the new compressed sentences. This finding was in line with the evaluation by the panel of experts; they rarely removed these types of terms in the manual summary construction and the removal procedure was not done based on the individual terms. In contrast, if the FASPe *conf* value was set higher, i.e., 0.7, the number of compressed sentences ws only 54, but only one (1) invalid compressed sentence was found.

Table 4

Total Invalid Sentence Rate by FASPe Conf Value for a Compressed Sentence.

FASPe *conf* value	Total Compressed Sentences	Total Invalid Sentences	(%)
1	7	0	0
0.9	19	0	0
0.8	47	0	0
0.7	54	1	1.85
0.6	109	5	4.59
0.5	161	9	5.59
0.4	333	31	9.31
0.3	521	44	8.45
0.2	551	54	9.8
0.1	553	56	10.13

Referring to Table 4, if the FASPe *conf* was set to 0.6, it was discovered that out of 109 sentences that were compressed, five compressed sentences were found invalid by the proposed sentence

validation technique. The invalid compressed sentences were the ones that did not belong to any of the Malay grammar patterns based on the new Malay phrase derivation sets using the word class technique presented earlier in Table 3. Next, Figure 6 illustrates an example of an invalid compressed sentence using the FASEe *conf* value of 0.3, found using the proposed technique. After the sentence segmentation step, the original sentence was divided into two segments. Based on the FASPe and FASP weightage calculation, the first segment was to be removed. In the sentence validation process, the new compressed sentence underwent the shallow POS tagging and was validated against the Malay grammar pattern. It was discovered that the compressed sentence now began with the term "*berteduh*". Since the term belonged to the KK word class, the new subject phrase was considered invalid where the new compressed sentence *cs* had a grammatical syntax error.

Figure 6

An Example of an Invalid Compressed Sentence Using FASEe Conf Value of 0.3 Found Using the Syntactic-Based Sentence Validation Technique

***Original sentence*:**
Mangsa yang kehilangan rumah pula, berteduh di bawah khemah yang didirikan di sebalik runtuhan kediaman.
Sentence Division by segments:
seg_1 : |Mangsa yang kehilangan rumah pula|
seg_2 :|berteduh di bawah khemah yang didirikan di sebalik runtuhan kediaman|

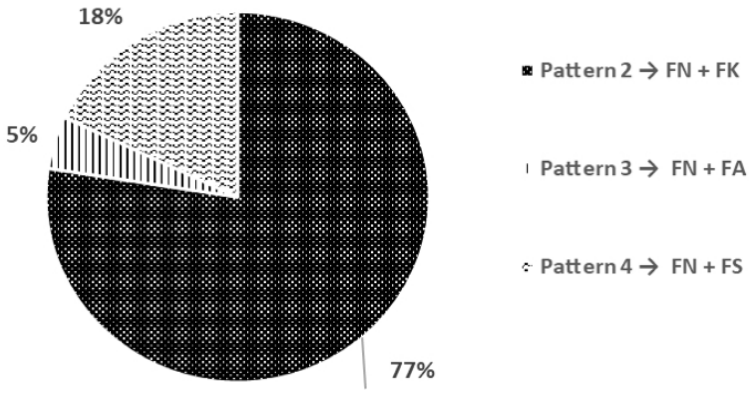
***Compressed sentence*:**
|~~Mangsa yang kehilangan rumah pula~~| seg_1
|berteduh/**KK** di/ktsn bawah/**KN** khemah/**KN** yang/**KT** didirikan/**KK** di/ktsn sebalik/**KA** runtuhan/**KN** kediaman/**KN**| seg_2
∴ Invalid cs : *Berteduh di bawah khemah yang didirikan di sebalik runtuhan kediaman.*

The findings are presented in Figure 7 in the form of a Malay grammar pattern that corresponded with the compressed summaries. Most of the valid compressed sentences fell in the Pattern 2 group, with 77 percent that consisted of the FN and FK phrases. This finding was considered relevant in the current study because most of the Malay sentence constructions belonged to this group. Therefore, a sentence that underwent the compression procedure should still preserve the noun and verb phrases that described the topic of the sentence.

The second highest group belonged to Pattern 3, which was 18 percent and the lowest belonged to Pattern 4 of FN + FS with only 5 percent. No compressed sentences belonging to Pattern 1 in this experiment were found to be built on top of FN + FN phrases. This finding was justified because complex Malay sentences were generated by more than one subject and predicate; thus, the FK phrase was more dominant to be found in a valid sentence.

Figure 7

Malay Grammar Pattern Group for Compressed Malay Sentences Based on the New Derived Syntactic-Based Rule



Automatic Evaluation

The ROUGE toolkit measures the quality of the automated summary by comparing it against the reference summaries created by human experts. It measures the overlapping agreement in chosen words such as the N-gram between the summarizer model and the ones produced by humans. The ROUGE-N is calculated using Equation 1, where the evaluation metrics produced are recall, precision, and F-measure.

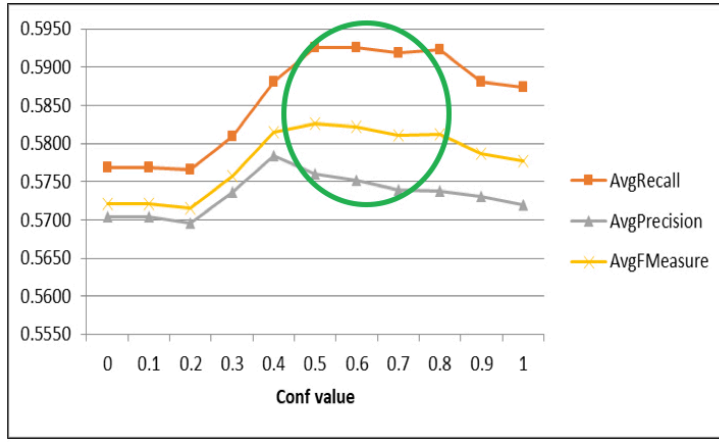
$$ROUGE - N = \frac{\sum_{S \in (ReferenceSummaries \quad \}gram_n \in S} \sum count_{match}(gram_n)}{\sum_{S \in (ReferenceSummaries \quad \}gram_n \in S} \sum count(gram_n)} \qquad (1)$$

The compressed summary produced by the MYTextSumCOMP method was evaluated against the manual summary produced by the Malay experts by using the ROUGE evaluation toolkit to find the optimum compression rate via the *conf* value.

By optimizing the *conf* value, the performance of the summarizer model was evaluated by using the ROUGE evaluation metrics. From Figure 8, it was found that better results were achieved as the *conf* value increased, whereby the *conf* values were around 0.5 and 0.6. However, it was also observed that when the *conf* value was set higher than 0.6, the summary's performance began to decrease. Referring to Table 4, it can be seen that the total percentage of invalid sentences for *conf* value of 0.5 was around 5.6 percent with 161 compressed sentences. With this setting, the highest average F-measure of 0.5826 and recall value of 0.5925 were found, which indicated the optimum 0.5 *conf* value for the summarizer model.

Figure 8

Performance Results of MYTextSumCOMP Based on Conf Value.



Manual Evaluation

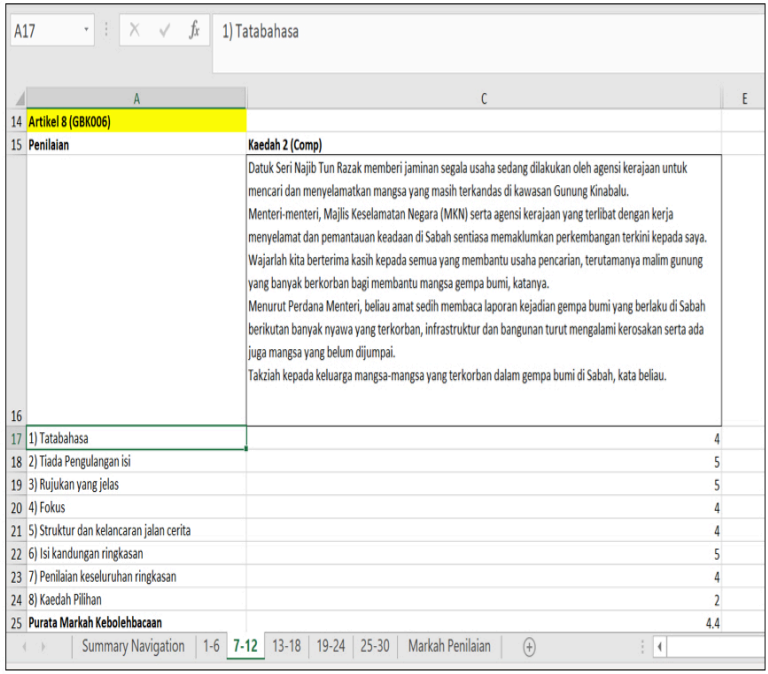
For the manual summary evaluation, this paper followed the DUC2005 guidelines, whereby 30 compressed summaries randomly produced by the MYTextSumCOMP model were given to three panels of experts. In this manual evaluation, the automated summaries were not compared to human summaries. The assessment by the panels was based on the content responsiveness and readability metrics, which included linguistics quality. The readability was based on linguistic quality metrics of grammaticality, non-redundancy, referential clarity, focus, and structural coherence.

The other evaluation was based on the summary's content responsiveness, which assessed the important content that was covered within the topic. The manual evaluation used a five-point Likert scale (i.e., 1: very bad, 2: bad, 3: average, 4: good, and 5: very good).

Figure 9 depicts the manual summary evaluation procedure, where each panel evaluated the given summary based on the content responsiveness and readability metrics. Since the compressed summary sentence was validated using the syntactic-based technique, the panel evaluated the overall grammar of the summary sentence by giving the score accordingly.

Figure 9

Evaluation Summary of the Manual Procedure by the Panel of Experts Following DUC2005 Guidelines.

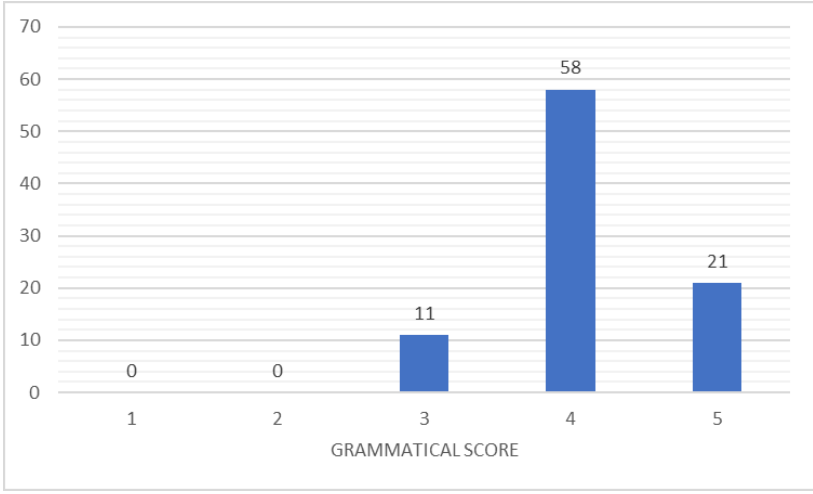


The detailed findings from the manual evaluation on the grammatical quality of the compressed summary sentence after the sentence validation produce are presented in Figure 10. It was found that out

of the 90 summaries (30 summaries evaluated by three panels), only 11 summaries were given an average score, with 58 good and 21 very good grammar quality. This indicated that th proposed sentence validation technique could filter invalid compressed sentences before being added to the final summary.

Figure 10

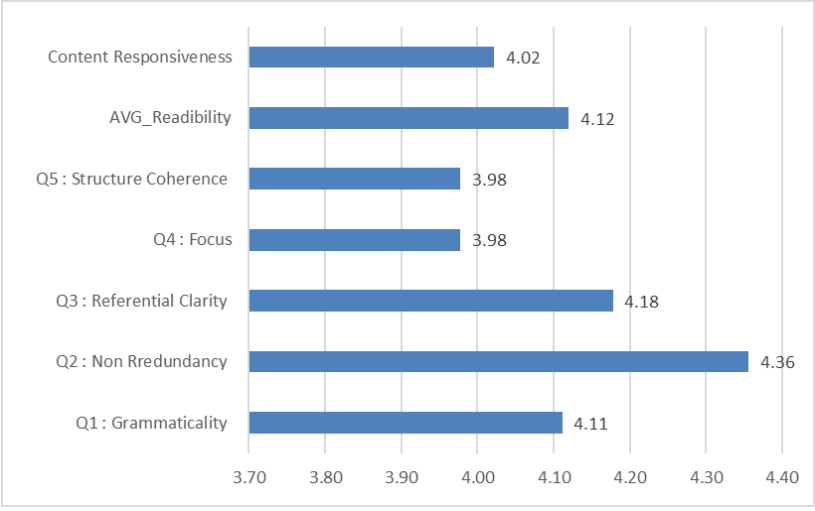
Detail Results on Manual Compressed Summary Evaluation by A Panel of Experts



Following this, Figure 11 summarizes the overall evaluation results, stating that the average grammatical score of the summary produced by the summarizer model was 4.11 out of 5. The average readability score was 4.12, whereas the content responsiveness score was 4.02. This indicated that the compressed summary produced was also readable and able to preserve the content without sacrificing the sentence's grammar.

Figure 11

Overall Results on Manual Compressed Summary Evaluation by A Panel of Experts.



CONCLUSION

This paper detailed the syntactic-based sentence validation technique implemented in the Malay text summarizer model named MYTextSumCOMP. The sentence validation technique consisted of two phases that are the shallow POS tagging by word class and the phrase structure and parsing. This study analyzed the Malay POS elements to construct a new derivation set of syntactic rules for Malay phrase validation based on Malay main word classes to validate grammatical Malay sentences. From the automatic summary evaluation, the highest average F-measure of 0.5826 and recall value of 0.5925 were found, which indicated optimum compression with a 0.5 *conf* value. Manual summary evaluation from a panel of experts also supported the findings, whereby the summary grammatical score achieved was good with a 4.11 out of 5 score. The proposed technique still manually refers to the Malay POS corpus, in which new Malay words need to be tagged to properly classify them into the respective Malay word classes. It is recommended for this work to be integrated with any existing automatic Malay parser to automate this process

as significant research in the Malay NLP area. Moving forward, the present researchers plan to test the proposed sentence validation technique using different datasets, such as social and technical datasets.

ACKNOWLEDGMENT

This work is supported by Universiti Sains Malaysia (USM), Research University Grant (RU) with the project number 1001/PKOMP/811295. The authors would also like to thank the panels of Malay language experts in providing the summary extract.

REFERENCES

Ab Aziz, M. J., Ahmad, F., Ghani, A. A. A., & Mahmod, R. (2006). Pola grammar technique for grammatical relation extraction in Malay language. *Malaysian Journal of Computer Science, 19*(1), 59.

Alfred, R., Mujat, A., & Obit, J. H. (2013). A ruled-based part of speech (RPOS) tagger for Malay text articles. In *Asian Conference on Intelligent Information and Database Systems* (pp. 50–59). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36543-0_6

Alias, S. (2018). *A pattern-growth sentence compression technique for Malay text summarizer.* Doctoral dissertation, Universiti Sains Malaysia.

Ariffin, S. N. A. N., & Tiun, S. (2018). Part-of-speech tagger for Malay social media texts. *GEMA Online® Journal of Language Studies, 18*(4). http://dx.doi.org/10.17576/gema-2018-1804-09

Conroy, J. M., Schlesinger, J. D., O'leary, D. P., & Goldstein, J. (2006, November). *Back to basics: CLASSY 2006.* In *Proceedings of DUC* (Vol. 6, No. 460, p. 460).

Hamza, M. A., Ab Aziz, M. J., & Omar, N. (2019). Identification of sentence context based on thematic role rules for Malay short essay assessment. *International Journal of Software Engineering and Computer Systems, 5*(2), 66–77.

Jing, H. (2000, April). Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on*

Applied Natural Language Processing (pp. 310–315). https://doi.org/10.3115/974147.974190

Jusoh, S., Masoud, A. M., & Alfawareh, H. M. (2011). Automated text summarization: Sentence refinement approach. In P. J. Snasel V., El-Qawasmeh E. (Ed.), *Digital Information Processing and Communications. Communications in Computer and Information Science*, 189, 207-218. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22410-2_18

Knowles, G., & Don, Z. M. (2003). Tagging a corpus of Malay texts, and coping with 'syntactic drift'. In *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 422–428).

Lin, C.-Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop* (pp. 74–81).

Mohamed, H., Omar, N., & Ab Aziz, M. J. (2011). Statistical Malay part-of-speech (POS) tagger using Hidden Markov approach. In *2011 International Conference on Semantic Technology and Information Retrieval* (pp. 231–236). https://doi.org/10.1109/STAIR.2011.5995794

Mohd Don, Z. (2010). Processing natural Malay texts: A data-driven approach. *Trames, 14*(1), 90–103. https://doi.org/10.3176/tr.2010.1.06

Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa & Mahmood, A. H. (2008). *Tatabahasa Dewan Edisi Ketiga*: Dewan Bahasa dan Pustaka.

Omar, A. H. (1998). *Morfologi-sintaksis bahasa Melayu (Malaya) dan bahasa Indonesia: Satu perbandingan pola*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Tran, N.-T., Ung, V.-G., Luong, A.-V., Nghiem, M.-Q., & Nguyen, N. L.-T. (2015). Improving Vietnamese sentence compression by segmenting meaning chunks. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference* (pp. 320–323). https://doi.org/10.1109/KSE.2015.74

Wang, L., Raghavan, H., Castelli, V., Florian, R., & Cardie, C. (2013). A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria*.

Xian, B. C. M., Lubani, M., Ping, L. K., Bouzekri, K., Mahmud, R., & Lukose, D. (2016). Benchmarking Mi-POS: Malay part-of-

speech tagger. *International Journal of Knowledge Engineering, 2*(3), 115–121. https://doi.org/10.18178/ijke.2016.2.3.064

Zajic, D., Dorr, B. J., Lin, J., & Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing & Management, 43*(6), 1549–1570. https://doi.org/10.1016/j.ipm.2007.01.016

Zamin, N., & Ghani, A. (2010). A hybrid approach for Malay text summarizer. In *Proceedings of the International Multi-Conference on Engineering and Technological Innovation*.

Zamin, N., Oxley, A., Abu Bakar, Z., & Farhan, S. A. (2012). A statistical dictionary-based word alignment algorithm: An unsupervised approach. In *2012 International Conference on Computer and Information Science, ICCIS 2012 - A Conference of World Engineering, Science and Technology Congress, ESTCON 2012 - Conference Proceedings, 1*, (pp. 396–402). https://doi.org/10.1109/ICCISci.2012.6297278

APPENDIX

Algorithm 1. *A syntactic-based sentence validation technique for Malay Text Summarizer using New Malay Phrase Derivation Sets.*

```
Algorithm 1 : msValidation(ms)
Input : A malay sentence ms  with POS tags
Output : A valid ms with Malay Grammar Pattern P class

//Initialization of Malay Grammar Pattern P
Pattern 1 → FN + FN
Pattern 2 → FN + FK
Pattern 3 → FN + FA
Pattern 4 → FN + FS

//New Malay phrases derivation set
FN → {KN,  KT}
FK → {KK,  KT,  KN(Object)}
FA → {KA,  KT}
FS → {KSN,  KT,  KN}

1:For each compressed sentence cs ∈ s do
2:  Split the cs into segments seg_n using KN as delimiters

//1) Subject Phrase Validation (S can only consist of FN)
3: For each seg_i ∈ cs  do
4:  For each token i ∈ seg_1 do
5:   If (i == 1 ⊆ FN)
6:       validSubject = true;
7:   If (i == 1 ⊆ (KK,KA))
8:       validSubject = false
9:      break
10: end

//2) Predicate Phrase Validation
11: While validSubject = true
12: Check next seg_i
13:      If (seg_i ⊆ FN)
14:          Pattern P = 1;
15:      else if (seg_i ⊆ FK)
16:          Pattern P = 2;
17:      else if (seg_i ⊆ FA)
18:          Pattern P = 3;
19:      else if (seg_i ⊆ FS)
20:          Pattern P = 4;
21:       else
22:          Invalid Pattern P
23: Return Pattern P
24: end

//3) Sentence Validation
25: If (validSubject == true && Pattern P is valid)
26:      Return a valid cs with Pattern P
27: If (validSubject == true && Pattern P is invalid)
28:      Invalid cs
29: Else
30:      Unclassified cs
```