IDENTIFICATION OF SENTENCE CONTEXT BASED ON THEMATIC ROLE RULES FOR MALAY SHORT ESSAY ASSESSMENT

Mohd Azwan Hamza[1], Mohd Juzaiddin Ab Aziz[2], Nazlia Omar[3]

[1]Faculty of Software Engineering & Computer Systems, Universiti Malaysia Pahang
[2,3]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia

[1]azwan@ump.edu.my, [2]juzaiddin@ukm.edu.my, [3]nazlia@ukm.edu.my

ABSTRACT

Integrated short essay assessment can benefit related parties in education domain. It can be performed using linguistic approach to recognize sentence structure (word order). Subject-predicate rule does not consider sentence usage context in a sentence. Whereas, by using thematic role rule, each significant argument in a sentence is discovered to provide information on relation between word and the role it played in determining the meaning of a sentence. Tokens that have been labelled were matched with thematic role consisting of Agent, Patient, Theme, Source, Beneficiary, Experiencer, Location, Time and Quantity. This study is conducted on Compiler course in Malay which comprise of various type of sentences. 5-fold cross validation test shows that error rate produces by Thematic Role rule and Subject-Predicate rule are 13.51% and 34.05%, respectively, a difference of 20.54%. This result shows that by understanding sentence context, it will produce better and promising essay assessment.

Keywords: Thematic role, subject, predicate.

INTRODUCTION

The presence of e-learning and digital learning even computer-based national final examination in the level of secondary schools requires schools and universities to develop smart education (Nur et al., 2018). Nevertheless, essays have been neglected in many computer-based assessment applications since there exist few techniques to score essays directly by computer (Foltz & Landauer, 1999).

Essay assessment is a vital task in evaluating the quality of essay writing on particular topic. The process of writing in essay format involves analytical and critical thinking, and communication skill, which are suggested for intense learning approach, which enable the student to choose, arrange and describe their knowledge and understanding by presenting meta theory knowledge and conditional knowledge if they possess it (Blood, 2011; Boulton-Lewis, 1995). Short essay is written in simple sentences where writing style is not emphasised for grading (Mohd Juzaiddin et al., 2009).

The main objective of this research is to study, analyze and develop a set of thematic role rule, as well as sentence normalization process on training level to indicate thematic role of each significance argument comprising of Agent, Patient, Theme, Source, Beneficiary, Experiencer, Location, Time and Quantity, which are influenced by several factors such as type of verb, morphology and other dominant factors. Comparison was made based on Subject-Predicate rule in identifying sentence context for short Malay essay.

BACKGROUND AND PREVIOUS RESEARCHES

Malay language is an official statute in Malaysia and it widely used also in Brunei, Indonesia, and Singapore (Wafaa et al., 2018). Although there are many researchers conducted to measure sentence similarity for English, only few researches focused on Malay. One of the main factors is lack of basic tool in processing Malay language through linguistic method (POS tagger, semantics, ontology). Shahrul Azman et al. (2007) specifically tried to measure Malay sentence similarity based on vector-based approach. It uses pre-process Malay dictionary and Overlap Edge Counting-based approach to calculate word-to-word semantic similarity at the preliminary stage. From the study, the final outcome is encouraging and consistent if the usage of the approach is compared with human assessment. However, several issues have arisen. Based on experiment at the early stage, two similar sentences that have been assessed by the system were only 57.9% similar after ignoring morphology element. In contrast, the similarity level increased to 90.2% if the element was considered. Thus, sentence similarity measurement using word-to– word technique should be improved by applying other methods, such as corpus-based word occurrence and semantic network that is capable of measuring sentence similarity based on the real context of the sentence.

Further research in measuring Malay sentence similarity have been done based on new linguistic method, namely Pola Grammar Technique. The method measures similarity using sentence Grammar Relationship (GR) comparison (Mohd Juzaiddin, 2008). At the beginning, it will extract GR in the sentence and later match it with other sentences. The similarity is determined based on whether all subjects, verbs and objects are similar. If it is found only the subject is similar, it will subsequently refer to synonym verb in Malay thesaurus. The results showed assessment difference averages between this method and human evaluation were as low as 0.049, 0.028 and 0.12 for three sets of questions. With 0.05 significance level, it can be concluded that this evaluation is close to human assessment. Nonetheless, Pola Grammar technique can be optimized by integrating lexical semantic based on argument rule on function word to acquire related domain size.

The effort was continued by Suhaimi et al. (2011) who applied rule-based phrase database and knowledge base of synonym context to obtain specific word that is synonym in Malay sentence. Malay POS tagger used to mark lexical type on each tokenized word in the inserted text. Then, dependence-based word classifier will identify pattern match between words that is labelled by rule in rule-based phrase database. In order to determine the context of sentence in Malay text, context-based similarity determination module is implemented. By applying head and modifier rules, it will determine the context of the sentence by conducting searches in synonym context knowledge base. Nevertheless, a more comprehensive approach in identifying the context of sentence in each phrase using head and modifier rules is much needed.

THE APPROACHES TO IDENTIFY SENTENCE CONTEXT

Statistical method does not always able to identify a perfect match without clear relations or concept between two natural sentences. Some approaches deal with this problem via determining the order of words and the evaluation of semantic vectors; however, they were hard to be applied to compare the sentences with complex syntax as well as long sentences and sentences with arbitrary patterns and grammars (Lee et al., 2014). Disambiguation requires an accurate quantifier that can measure the semantic relation between any two sense (Al-Saffar et al., 2018).

In order to overcome this problem, researchers used semantic method (Lee, 2012; Mandreoli et al., 2002). This method uses semantic network, such as Wordnet, Vector Space Model and Statistical Corpus to calculate semantic similarity between words using different measurement methods. However, the semantic method measures similarity sentence only based on semantic similarity between words, where syntax information and other semantic knowledge, such as semantic class and thematic role are ignored (Wafa Wali et al., 2017).

Indeed, several knowledge features are not considered in sentence similarity measurement such as thematic role, semantic class and relation between syntax and semantic levels based on semantic predicate (Wafa Wali et al., 2017). When two sentences have similar syntax structure (subject + verb + object) but different verb semantic class, the sentences pair is similar syntactically, whereas actually both sentences are different according to human expert. For example, the sentence 'Ahmad kicks the ball' and the sentence 'Ahmad has a ball', have similar syntax structure and semantic relation for each argument (subject (noun) + verb + object (noun)) that exists in both sentences, but differs from the aspect of thematic role played by verb, consequently causes both sentences become dissimilar entirely. Thus, this study attempts to apply thematic role rules in order to determine specific sentence context so that students' answer and answer scheme will be compared syntactically similar.

THEMATIC ROLE RULES

Sentence context could not be merely measured based on word order or syntax structure. Therefore, this research uses linguistic approach, that is thematic role in discovering particular thematic role played by argument in a sentence, subsequently identifying sentence context. Semantic class and thematic role for each argument in a sentence could provide information regarding relation between words and the role played in determining meaning of a sentence (Wafa Wali et al., 2017). The function of thematic role rule is to identify the role played by verb in Malay thematic labelling (Ramli, 2006).

1.   Selection of Syntax Category

The selection of syntax category refers to a method to identify chosen category for a particular verb. It determines number of Name Phrase (NP) that needs to be exist depending on the Verb Phrase (VP). Verb can be divided to three categories: Transitive Verb (TV), Dual Transitive Verb (DV) and Intransitive Verb (ITV). For VP that contains TV, such as word hurai (parse), it needs a direct object NP as a recipient. Thus, the verb needs a complement. For VP that has DV, also known as compound transitive verb, such as tukar (change), it needs two recipients, namely two direct objects NP or one direct object NP and one indirect object Phrase Complement (PC). Thus, two complements are required. In contrast, VP that contains ITV, such as jalan (walk), does not require any complement. For proper noun, such as word ialah, adalah and merupakan (is), a direct object NP is needed as the recipient.

(i)   *Penghurai menghurai token* (Parser parse the token).
(ii)   *Nahu bebas konteks menukar token kepada frasa nahu* (Context-free grammar changes the token into grammar phrase).
(iii)   *Pemproses sedang berjalan* (Processor is running).
(iv)   *Penganalisis leksikal adalah komponen pengkompil* (Lexical analyzer is a compiler component).

Sentences (i) to (iv) are examples of sentences that use either TV, DTV or ITV. The categories are determined based on types of stem. Based on Chomsky's theory (1981), this type of verb is described in distribution framework or subcategory framework. According to

Ramli (2006), the distribution framework form can be converted into argument structure form that represents type of subject, verb and predicate.

(i)   *hurai* (parse): verb   1   2
                                    NP   NP
(ii)   *tukar* (change): verb,   1   2   3
                                    NP   NP   NP/CP
(iii)   *jalan* (walk): verb,   1
                                    NP
(iv)   *ialah/adalah/merupakan* (is): verb,   1   2
                                                NP   NP

2.   Thematic Role

Theta role or thematic role refers to semantic relations between verb and its argument (Ramli, 2006). For example, verb *menghurai* (parse) needs two arguments, which are labelled by its thematic role. Subject argument is labelled as Agent, while object is labelled as Patient.

3.   Type of Verb in Labelling the Thematic Role

Ramli (2006) has set seven significance thematic roles which implicates by verb: Agent, Patient, Theme, Source, Location, Benefiter and Experiencer. Meanwhile, Sri Liyaningsih & Siti Zuhriah (2016) have stated 12 thematic roles with addition of Tool, Quantity, Goal, Reason and Time). However, the number and type of selected thematic roles are based on training and testing data sets of a research.

The entire thematic roles can be described as these sentences:

(i)   *Penganalisis sintaksis menjelajah token* (Syntax analyzer explores the token).
                                                [Agent]   [verb] [Patient]
(ii)   *Pengkompil merupakan penterjemah* (Compiler is a translator).
                                        [Theme][verb][Theme]
(iii)   Nahu bebas konteks mempunyai terminal sebagai perwakilan token dalam Bahasa.
        (Context-free grammar has terminal as token representation in language)
                [Theme]   [verb][Theme]   [Location]

Thematic role labelling is occurred at base structure level. This means the role is permanent even though active sentence is modified to passive sentence. For example, NP Penganalisis Sintaks (Syntax Analyzer) in sentence (i) will remain as Agent although it is modified into Token dijelajah oleh penganalisis sintaksis (Token [Patient] is explored [verb] by syntax analyzer [Agent]).

4.   Morphology in Labelling the Thematic Role

Sentences (i) to (iii) are examples on how thematic role is labelled based on type of verb either transitive, dual transitive and intransitive. Nevertheless, in Malay, other than type of verb, morphological elements also influence the thematic role labelling.
For example:

(i)   *hurai* (parse)
        a. *Penghurai menghurai ayat* (Parser parses the sentence).
        b. *Penghurai menghuraikan ayat kepada token* (Parser parses the sentence into token).

69

(ii)   *lahu* (idle)
   a. *Pemproses sedang melahu* (Processor is idling).
   b. *Aturcara melahukan pengkompil itu* (Program idles the compiler).
   c. *Aturcara melahukan pengkompil itu di peringkat proses* (Program idles the compiler at process level).

Example (i) and (ii) shows different argument structure for verb *hurai* and *lahu*. The differences are influenced by prefixes and suffixes in the verb (denotes as underline).

(i) *hurai*
   a. <u>*meng*</u>*hurai* (parses):   verb; 1   2
                                                    NP   NP
   b. <u>*meng*</u>*hurai*<u>*kan*</u> (parses): verb; 1   2   3
                                                    NP   NP   NP
(ii) *lahu*
   a. *melahu* (idling):   verb; 1
                                    NP
   b. *melahukan* (idles):   verb; 1   2
                                                    NP   NP
   c. *melahukan* (idles):   verb; 1   2   (3)
                                                    NP   NP   NP

Verb *hurai* is a TV while verb *lahu* is an ITV. However, the change in argument structure is influenced by stem on the verb. Indirectly, the stem gives significance effect to thematic role played by verb in a sentence.

(i)   *hurai*
   a. *Penghurai menghurai ayat* (Parser parses the sentence).
                    [Agent] [verb]   [Patient]

   b. *Penghurai menghuraikan ayat kepada token*
       (Parser parses the sentence into token).
       [Agent][verb]   [Patient] [Beneficiary]
(ii)   *lahu*
   a. *Pemproses sedang melahu* (Processor is idling).
                    [Experiencer]
   b. *Aturcara melahukan pengkompil itu* (Program idles the compiler).
                        [Agent] [verb] [Patient]
   c. *Aturcara melahukan pengkompil itu di peringkat proses*.
       (Program idles the compiler at process level).
          [Agent] [verb] [Patient]   [Location]

From the discussion, it clearly shows Malay thematic role not only influenced by types of verb itself (either transitive or intransitive); however, the addition of prefixes and suffixes also provide a significance effect on sentence argument structure and affect Malay thematic role rule, indirectly.
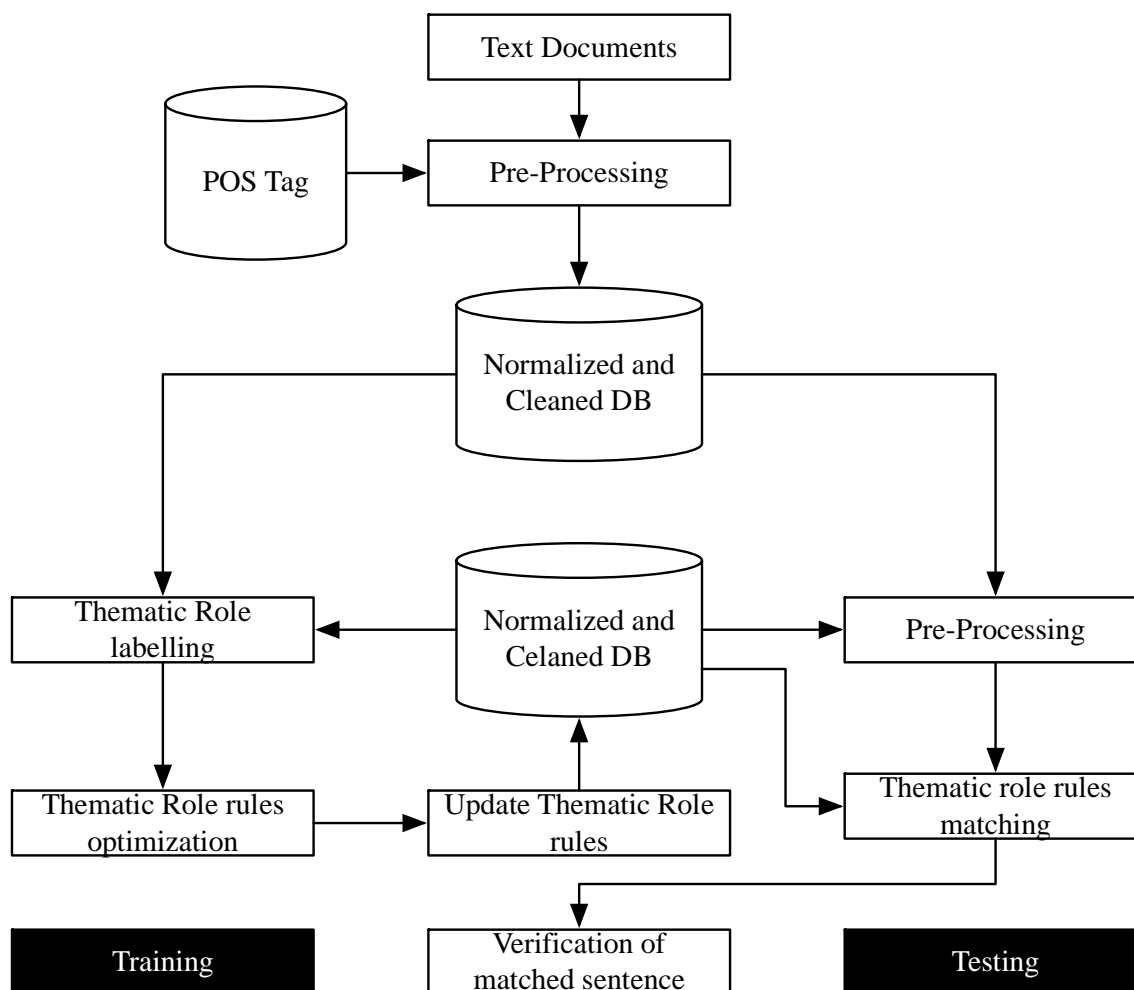
RESEARCH DEVELOPMENT METHODOLOGY

Figure 1. The architecture of the determines and labelling thematic role in Malay sentence.

Pre-Processing

Based on Figure 1, it shows the process to determines and labelling specific thematic role of significance arguments in Malay sentence. The process started with text documents as an input data set consists of question, answer scheme and students' answer question documents. Then, it involves pre-processing to prepare the data with proper form before it enters the main process. Pre-processing consists of four elements; tokenization, part-of-speech (POS) tagging, normalization and cleaning.

In tokenization, sentence will be chunked into smaller parts consist of words and symbols. Then, it will be merged to construct certain phrases based on two rules; proper noun and compound word. Then, POS tagger will tag all particular tokens. Once each token is tagged, compound phrase that are exists in the sentence will be identified. For proper noun rule, the matching formed is determined by first capital letter on each consecutive word. Whereas, for compound word rule, it is consisting of four type of compound words: compound nouns, compound verbs, compound adjectives and compound function words. After tagging on words and phrases are completed, sentences from students' answer document will be normalized. Sentence normalization aims to complete imperfect sentence from the aspect of construction structure of standard Malay sentence. Sentence normalization

process involves substitution of pronoun with predicate from previous question or sentence, restructuring of subject and predicate for sentence started with verb, fragmentation of sentence which is connected by conjunction, fragmentation of sentence which is connected with full stop and comma, replacement of verb *melakukan* (do) with prefix insertion '*meN*' + verb and discarding the word *ialah*, *adalah* and *merupakan* (is) if there are a more significance verb presence in a sentence. The last process in this phase is sentence cleaning. Sentences in all documents are cleaned from all stop words and symbols. This process is vital to ensure the existence of irrelevant tokens will not affect negatively the training and testing processes.

Training

In each *k*th-fold, four Data Sets were used for training. The training process are purposely to gain an optimized set of thematic role rules which will compromise few factors in labelling all significance arguments in a sentence; type of verb, morphological elements and any presence of special words. It starts with labelling the role of sentence from normalized and cleaned database with fundamental rules available in rules of thematic role database. Further, the labelled sentence will be verified and optimized based on previous three factors. Any changes made, will be update and save in the rules of thematic role database. This iteration process will continue until all sentences in the particular data set have been processed. Table 1 shows the final set of thematic role rules successfully generate by this research.

Table 1. Full set of thematic role rules.

Morphology	Special Word	Rules of Thematic Roles
meN..., meN-...-kan, memper-...-i, memper-...-kan	-	Agent-v-Patient
	kepada	Agent-v-Patient-(*kepada*)Theme/Beneficiary
	iaitu	Agent-v-Patient-(*iaitu*)Theme
	daripada, kepada	Agent-v-Patient-(*daripada*)Source-(*kepada*)Theme/Beneficiary
	dari-ke	Agent-v-Patient-(*dari*)Source-(*ke*)Theme/Beneficiary
	kepada, dalam	Agent-v-Patient-(*kepada*)Theme/Beneficiary-(*dalam*)Location
	kepada	Agent-v-Patient-(*kepada*)Theme/Beneficiary
	dalam	Agent-v-Patient-(*dalam*)Location
	pada	Agent-v-Patient-(*pada*)Location
	semasa	Agent-v-Patient-(*semasa*)Time
me-...-i	-	Agent-v-Experiencer
di-...	*oleh*	Patient-v-Agent
	ke/kepada	Patient-v-(*ke*)Theme/Beneficiary
	semasa	Patient-v-(*semasa*)Time
	oleh, dalam bentuk	Patient-v-Agent-(*dalam bentuk*)Theme
di-...kan	*mengikut/berdasarkan*	Patient -v-(*mengikut*)Theme
	semasa	Patient -v-(*semasa*)Time
	daripada	Patient -v-(*daripada*)Source
	untuk	Patient -v-(*untuk*)Beneficiary
	sebagai	Patient -v-(*sebagai*)Theme
	dalam	Patient -v-(*dalam*)Location
menerima, mendapat, memperoleh	*dalam bentuk*	Beneficiary -v-Theme-(*dalam bentuk*)Theme
menjadi	-	Experiencer-v-Theme
mempunyai, terdapat	-	Beneficiary -v-Theme
	dalam	Beneficiary -v-Theme-(*dalam*)Location
iaitu, bertindak sebagai	-	Theme-v-Theme
	kepada, dalam	Theme-v-Theme-(*kepada*), Beneficiary/Theme-(*dalam*)Location
merupakan, adalah, ialah, berupa, terdiri daripada	-	Theme-v-Theme
	kata bilangan	Theme-v-Theme-(*Quantity*)
	dalam	Theme-(*dalam*)Location-v-Theme
	bagi	Theme-v-Theme-(*bagi*)Beneficiary
digunakan	*untuk*	Patient-v-(*untuk*)Beneficiary
	dalam	Theme-v-(*dalam*)Location
	pada	Theme-v-(*pada*)Location
	oleh	Patient-v-(*oleh*)Agent

Testing

In order to prove the effectiveness of the thematic role rules which were constructed earlier, one data set for each cycle of kth-fold is utilized for testing. Sentence in testing data set from normalized and cleaned database, once again, is taken as an input. There are only two steps in this testing process which are labelling and matching. Particular sentence will be labelled

with a specific thematic role by referring to the rules of thematic role database. After that, the matching is done by comparing pattern of the rules in student's answer and answer scheme.

Output

As a final outcome, it will verify whether particular sentence is matched with pattern of rules existed in any answer scheme (if there are more than one sentence). For the purpose of comparison, the outcome will compare with the one that uses subject-predicate rules.

RESULT

The study shows the effectiveness of thematic role rule implementation on sentence structure compared to general sentence structure built based on subject-predicate rule. In this research, *k*-fold cross validation method was used.

Simple sentence is easier to process than compound sentence. Simple sentence is built from one subject and one object depending on types of VP either TV or ITV. Whereas compound sentence can be extremely hard to identify the accurate sentence structure pattern because it is developed from combination of more than one verb phrases and objects. It becomes more complex when involving preposition and combination of singular sentence and compound sentence. For this reason, the entire students' answer sets were divided equally. There are 67 students' answer sets in Part A, 41 in Part B, 34 in Part C and 43 in Part D, separated into five sets namely Data Set A, B, C, D and E, representing 5-fold cross validation. Each data set has 37 students' answer sets that was equal in terms of quantity and combination of various types of simple and compound sentences so that the results are fair enough. There were 5-fold cross validation where each kth-fold, different data set from the entire Data Set A, B, C, D and E will take turn to become testing data set while the others become training data set.

Sentence similarity is measured by comparing the effectiveness of argument structure matching labelled by thematic role rules against argument structure matching labelled by subject-predicate rules and synset matching on semantic relation based on Wordnet taxonomy.
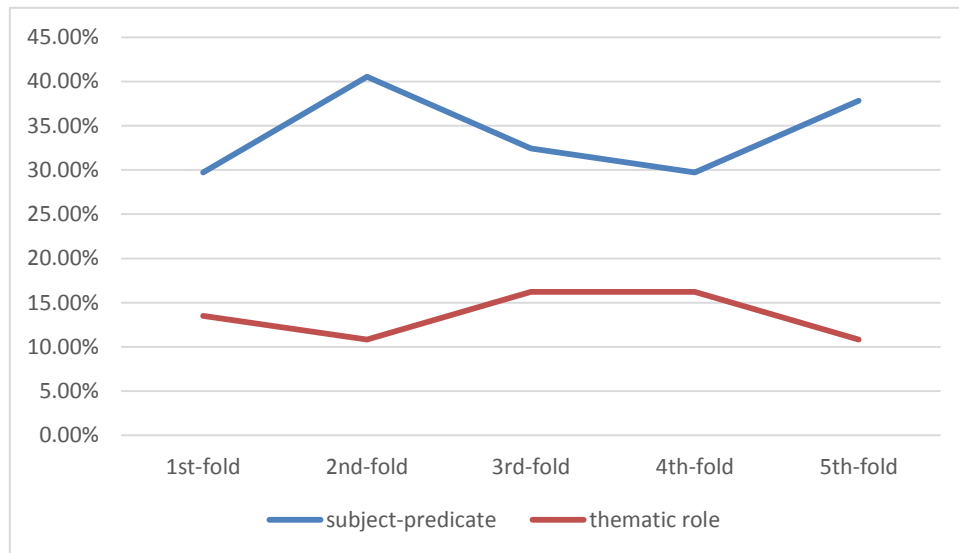

Figure 2. The Percentage of Error Rate of Answer Set Matching with SP and PT rules.

The implementation of thematic role rule compared to subject-predicate rule demonstrates a rather high error rate of similarity measurement. Based on Figure 2, the error percentage of answer set matching using subject-predicate rule was around 33% compared to only 14% for thematic role rule.

Table 2. The Average of Effectiveness Test.

	Subject-Predicate Rules	Thematic Role Rules
Minor Error	16.75%	3.78%
Major Error	17.30%	9.73%
Total of Error	34.05%	13.51%
Error Difference	20.54%	

On average, as shown in Table 2, based on the 5-fold cross validation tests, overall error (major and minor) between thematic role rules and subject-predicate rules was 13.51% and 34.05% respectively, a difference of 20.54%. The statistical findings of the research were contributed by several factors, which are:

(i)   The absence of subject or predicate in a sentence. When this happens, sentence structure matching based on subject-predicate rule fail to be obtained because it requires the presence of subject+verb+object pattern. In contrast, for thematic role rule, sentence is labelled based on the existence and types of verb and NP (object) in the sentence.

(ii)   Subject-predicate rule implementation only do the match of sentence structure pattern of subject + predicate (verb and object). This rule does not consider the context of words usage in a sentence.

Succinctly, although with sentence structure that is less than perfect but still grammatical correct, the application of thematic role rules still able to match sentence accurately. Indeed, these rules can identify sentence context by marking thematic role for each significance argument in a sentence. Without this control, the probability of error in synset similarity measurement to occur is higher.

CONCLUSION

The research was done based on two main objectives, namely to study, identify and develop an optimize set of thematic role rules and implement it to validate sentence structure in students' answer set. Finally, achievement verification was applied to prove the effectiveness of the rules set that has been established.

The research on thematic role was started with the research outcome by referring to the main research by Ramli (2006) and Seong (2011) that outlined the determination of thematic role was not just by verb, but stemming also played a vital role. However, the outcome of this research discovered one more factor that affected the development of thematic role rule, which was the existence of several special words. The findings from this research also discovered nine types of thematic role involved: Agent, Patient, Theme, Source, Beneficiary, Experiencer, Location, Time and Quantity.

Nevertheless, just as the objectives been set at the beginning of research, the implementation of training process on related data set was hard to be performed without

several normalization process that gave direct effect on the establishment of thematic role rule set. The set of established thematic role rule was constructed to mark thematic role for each related argument (NP subject, verb and NP object) at the training stage of the research. The number and types of significant arguments are based on rule set that has been constructed. Thematic role labelling is intended to identify sentence structure in test set, scheme set and students' answer set. Thus, sentence context based on the labelled argument thematic role can be determined. Next, 5-Fold Cross Validation was applied to substantiate the implication of the thematic role rules usage compared to subject and predicate rules. The outcome of the effectiveness test demonstrated that overall error (summation of minor and major errors) are 13.51% and 34.05% respectively, a difference of 20.54%.

REFERENCES

Nur, M. N. A., Musaruddin, M., Bunyamin, & Zulkaida, W. O. (2018). Concept of Smart City for Education: A Case Study in Kendari, Southeast Sulawesi, KnE Social Sciences, The 2nd International Conference on Vocational Higher Education (ICVHE) 2017, 1558–1565.

Foltz, P., Darrell, L., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic *Journal of Computer - Enhanced Learning*.

Blood, I. (2011). An Automated Essay Scoring: A Literature Review. Published in partnership with Columbia University Libraries and Applied Linguistic and TESOL at Teachers College 11, 40-64.

Boulton-Lewis, G.M. 1995. The SOLO Taxonomy as a Means of Shaping and Assessing Learning in Higher Education. Higher Education Research and Development 14: 143-154.

Mohd Juzaiddin Ab Aziz, Fatimah Dato' Ahmad, Abdul Azim Abdul Ghani & Ramlan Mahmod. (2009). Automated Marking System for Short Answer Examination. IEEE Symposium on Industrial Electronics and Applications, 47-51.

Shahrul Azman Noah, Amru Yusrin Amruddin & Nazlia Omar. (2007). Semantic Similarity Measures for Malay Sentences. *10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers*, 117-126.

Mohd Juzaiddin Ab Aziz. (2008). *Pola Grammar for Automated Short Answer Essay-Typed Examination*. Doctoral Thesis, Universiti Putra Malaysia.

Suhaimi Ab Rahman, Nazlia Omar, Hassan Mohamed & Mohd Juzaiddin Ab Aziz. (2011). A Synonym Contextual-based Process for Handling Word Similarity in Malay Structure. International Conference on Advanced Science, Engineering and Information Technology, 248-252.

Lee, M., Chang, J. W., & Hsieh, T. (2014). A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences. *The Scientific World Journal*.

Lee, M. C., Chang, J. W., Hsieh, T. C., Chen, H. H., Chen, C. H. (2012). Similarity Measure based on Semantic Patterns. International Conference on Advanced Information Science and System (AISS) 4(18): 10.

Mandreoli, F., Martoglia, R., & Tiberio, P. (2002). A Syntactic Approach for Searching Similarities within Sentences. Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02, 635-637.

Wafa Wali, Bilel Gargouri1 & Abdelmajid Ben Hamadou. (2017). Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge. *Vietnam Journal of Computer Science 4*(1), 51-60.

Ramli Md. Salleh. (2006). Imbuhan dan Penandaan Tematik dalam Bahasa Melayu. *Jurnal Melayu 2*, 19-51.

Sri Liyaningsih & Siti Zuhriah Ariatmi. (2016). *A Thematic Roles Analysis of Simple Sentences Found on The Titles of China Daily Newspaper*. Thesis, Universitas Muhammadiyah Surakarta.

Seong, G. S.: Penterjemahan Kata Kerja Bahasa Cina-Bahasa Melayu: Satu Analisis Ketepatan Makna Padanan. In: Journal of Language Studies, vol. 11(1) (2011)

Al-Saffar, A., Awang, S., Tao, H., Omar, N., Al-Saiagh, W., & Al-bared, M. (2018). Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. PloS one, 13(4).

Wafaa, A. S., Tiun, S., Ahmed, A. S., Awang, S., & Al-Khaleefa, A. S. (2018). Word sense disambiguation using hybrid swarm intelligence approach. PloS one, 13(12).