International Journal of Software Engineering and Computer Systems (IJSECS)
ISSN: 2289-8522, Volume 5 Issue 2, pp. 37-65, August 2019
©Universiti Malaysia Pahang
https://doi.org/10.15282/ijsecs.5.2.2019.4.0060

SENTENCE SIMILARITY MEASUREMENT BASED ON THEMATIC ROLE AND SEMANTIC NETWORK TECHNIQUES

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

Automated Short Essay Assessment is a subjective assessment that emphasizes important contents more than writing style. Word Order Technique and Syntactic-Semantic Knowledge Technique have been used in previous researches. However, it cannot differentiate sentence pair that is not similar semantically and only proven to produce excellent result on short sentence. Thematic Role annotation for every significant argument seems able to provide information regarding the relations between the word. Wordnet Semantics Network calculates semantic similarities of two *synsets* (token) by taking into account the depth of semantic relations. This study is conducted on Compiler course set in Malay that comprise of passive, simple, negative, mixed and complex questions. To prove the effectiveness of the techniques, the test result that apply Grammar Pola Technique was used as a benchmark because the study uses the same data set. The average of f-measure test accuracy rate is 93.53% when using Thematic Role and Semantic Network Techniques compared to 82.36% accuracy rate when using Pola Grammar Technique. The result of Thematic Role can be used on research involving Malay linguistic to test sentence structure matching that has verb by considering types of sentence.

Keywords: Thematic role, semantic network, *synset*, pola grammar.

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 et al., 1999).

Essay is important in assessing academic excellence by measuring students' ability to connect different ideas (Ramalingam et al., 2018). Essay assessment questions can be divided into two types: long essay and short essay. Assessment on long essay usually consists of grammar, usage, mechanics (spelling, punctuation marks, capital letters and paragraph) and style (Attali & Burstein, 2006). Meanwhile, short essay is written in short sentences and writing style is not emphasized for grading (Mohd Juzaiddin et al., 2008). Considering the limited number of words in short essay, every word or phrase in each sentence is significant in contributing some marks to the essay assessment. Integrated Essay Assessment (IEA) if proven effective not only will reduce assessment time but the comparison with human assessment will also produce a level that is almost similar (Ramalingam et al., 2018).

IEA is defined as computer technology that has the ability to evaluate the actual quality of a writing (Shermis & Burstein, 2016). The main focus of IEA system is marks generated by the system should be acceptable which is close to human assessment. To achieve the objective, this study applies the approach of sentence similarity measurement. By comparing students' answer documents and answer schemes, assessment is made based on the similarity level between these two text documents. The text documents are compared to obtain the accurate and acceptable marks based on the sentence similarity method (Dikli 2006; Erikson 2000; Fitzgerald 1994; Mohd Juzaiddin 2008). Assessment based on sentence similarity becomes complex when it involves more than one sentence structure. The degree of complexity is also influenced by several factors: the word position in the sentence structure and the ambiguity of word in the sentence.

The main objective of the study is to develop and apply a Rules of Thematic Role and sentence normalization process on training and test levels to annotate Thematic Role of each significance arguments, which comprises of Agent, Patient, Theme, Source, Beneficiary, Experiencer, Time and Quantity. The Thematic Role is influenced by several factors such as type of verb, morphology elements (prefix and postfix) and other dominant factors. For the purpose of semantic relations similarity measures between two tokens (*synsets*) and sentence similarity, modified Wu & Palmer's (wup) measurement method in Malay semantic relations based on *Wordnet* architecture is used.

BACKGROUND AND PREVIOUS RESEARCHES

In 1966, Ellis Page, the pioneer of IEA had started research by designing a computer program named *Project Essay Grade* (PEG) (Wang & Brown, 2007). With statistical capability constraint at that time, PEG attempted to determine the optimum combination of features weightage to make the best prediction that is close to human assessment using Multiple Linear Regression. In other words, PEG emphasizes essay assessment based on fundamental writing quality without considering content. The design of PEG approach is based on proxes, which comprises of length of essay, number of prepositions, relative pronoun and other part-of-speech (POS), as an indicator of the complexity level of sentence structure (Valenti et al., 2003).

Around 1980s, assessment pattern has evolved from merely assessing essay to provide feedback to students and teachers. Writer's Workbench (WWB) system developed by American Telephone & Telegraph had been designed to provide feedback to writers from the aspect of spelling, pronunciation and readability (Hearst 2000). WWB assess several features regarding style, average word length, sentence length division, types of grammar in a sentence, percentage of passive verb, and percentage of noun that has been normalized (Burstein & Wolska, 2003). The feedback received from this system can be used to enhance students' writing skill in the future.

The development of Natural Language Processing (NLP) and Information Extraction (IE) on late 1990s has open a new chapter with the production of more advanced English IEA designs. At that time, three main IEAs have been developed: e-rater, Intelligent Essay Assessor and IntelliMetric. By applying Step-Wise Linear Regression to determine assessment model that able to make the best prediction which is close to human assessment, e-rater evaluates essay based on variety of syntaxes, topic content and idea arrangement (Attali & Burstein, 2006).

By applying the approach of Latent Semantic Analysis, Landauer & Laham (2000) have developed Intelligent Essay Assessor that emphasises feedback preparation, which considers three elements: content, style and mechanics. Subsequently, Vantage Learning researchers reported that they have adapted Artificial Intelligent with NLP and statistical technology in developing Intellimetric and with this, it can analyse more than 300 features on semantic, syntax and argument levels (Valenti et al., 2003).

During the researchers' enthusiasm in extracting the most significant features on the assessment of long essay writing, some of the researchers started to study Integrated

Short Essay Assessment (ISEA). Most of the suitable techniques to be applied on long essay may not be suitable anymore to be implemented on short essay. This is because long essay usually has high frequency rate of word occurrence (Li et al., 2006). In short essay, every word presence has a high probability to be significant in contributing to certain marks. Short essay grading system is designed for short and factual answer, where the right or wrong criteria is clear (Raheel Siddiqi, 2010). The marks given are more reflected by content rather than writing style.

Currently, several systems have been successfully developed for English ISEA. Oxford-UCLESS system applies pattern approach, which is started with word and synonym set, then, makes particular search in essay to produce new pattern (Mohler & Mihalcea, 2009). IE method has been adapted in the system that processes ungrammatical and incomplete sentences in most UCLES case exam. Meanwhile, c-rater is an assessment engine of short essay, developed by ETS Technologies, which has been designed to give feedback on the answers obtained for the questions assessed. The system is capable to measure comprehension of sentence content. It uses pronominal reference of predicate phase structure, morphology analysis and synonym to evaluate the whole or parts of short essay question (Leacock & Chodorow, 2003). Apart from that, Automark has been developed for robust automated assessment on short free-text. IE method is used to extract concept or meaning from the free-text and the developer has striven as best as it could to ensure the system is able to tolerate well with typing error, spelling error, syntax errors and other errors (Raheel Siddiqi, 2010). It gives emphasis on content analysis without ignoring writing style features.

THE ASSESSMENT APPROACHES OF AUTOMATED SHORT ESSAY

One of the earliest researches in measuring sentence similarity was word presence frequency-based approach, also known as Bag-of-Words Model. The approach usually used in assessing open-text question, while other methods are more suitable for usage on closed-text question. Open-text question needs the students to write longer on topic asked with only be based on their experience and knowledge (Diana Perez, 2004). Meanwhile, closed-text question is divided into two: comprehensive question and structure question. For structure question, answers constructed could not be more than five sentences (Pulman & Sukkarieh, 2005).

Probability Model, Vector Space Model (VSM), Transformed Distance and N-gram Model are some of the usual techniques used in word occurrence frequency and word statistics (Shan et al., 2009). These techniques are based on the assumption that a document is considered as similar if the documents have many identical words (Ning et al., 2011). VSM is based on Term Frequency–Inverse Document Frequency (TFIDF). Even though TFIDF is a very effective model by adapting statistical-based technique, it is not yet appropriate for short text (Shan et al., 2009). This is because the frequency of identical word presence in two short text, which are similar, is low and perhaps non-existent, whereas both texts is equivalent because of the usage of different words, even though they carry the same meaning semantically.

Corpus-based statistical approach usually calculates sentence similarity based on statistical information using large size corpus. The technique is frequently used in Latent Semantic Analysis (LSA) model (Higgins & Burstein, 2007) and Hyperspace Analogues to Language (HAL) (Burgess et al. 1998). LSA involves the usage of Singular Value Decomposition (SVD) on document-term matrix in its effort to reduce rank. Several of LSA weaknesses: [1] word dimension size based on content matrix is limited due to the constraint existed in SVD calculation limit, subsequently leads to the occurrence of several important words from the input text (sentence) perhaps are not inserted into LSA dimension space and it will ignore syntaxes from the two sentences (Li et al., 2006). Regarding HAL, information on word presence frequency is also used to produce high space dimension. Nevertheless, result obtained shows HAL capability is not as good as LSA in measuring short text similarity. While, word-to-word matrix does not depict the actual meaning of a sentence (Ning et al., 2011).

As a comparison, knowledge-based semantic approach measures the similarity between two sentences based on semantic information extracted from the database (Ho et al., 2010). Normally, the semantic information is obtained by taking into account the closest meaning from the comparison of the two sentences. Ho et al. (2010) have developed a more optimum method: two sentences comparison based on actual meaning by modifying sentence similarity measures based on existing corpus to knowledge-based method. Li et al. (2006) measured semantic similarity between sentences or texts based on semantic relations and word order information. Semantic equivalence is obtained from knowledge database and Wordnet corpus. Subsequently, they considered word order implication on sentence meaning.

Vector-based semantic approach is usually used in IE system, where the most relevant document with input text determined by representing a document as word vector and input text is matched with the similar document in document database through similarity matrix (Islam & Inkpen, 2008). Among advanced research for vector-based technique is by using Random Indexing (RI) to seek document in semantic space (Higgins & Burstein, 2007). RI produces semantic vector for every word in the corpus, where it subsequently is compared with vector for the other word using cosines similarity standard matrix (Lee, 2010). Besides that, other advanced research in vector-based method that leads to a word-similarity sentence-based plagiarism detection (SimPaD) that applies sentence-to-sentence comparison (Gustafson, 2008). The technique is based on pre-count word correlation factor to identify sentence-to-sentence similarity and finally the similarity rate for any two documents is detected as plagiarism. Nonetheless, as SimPaD does not consider word order in sentence, thus, short text similarity measures perhaps are less precise.

Several other latest researches also show that the usage of semantic-based sentence measurement approach is able to achieve a good result. Wang et al. (2016) suggest a model that considers both similarity and dissimilarity by parsing and composing lexical semantics on sentence. The model represents every word as a vector and calculate semantic matching vector for every word based on every word in other sentences. Consequently, each word vector is parsed to similar component and dissimilar component based on semantic matching vector. Then, two-channel Convolutional Neural Network model is used to acquire features by parsing similar and dissimilar components. Finally, similarity measurement is calculated based on parsed features vector. The experiment result shows the achievement is equivalent with state-of-the-art achievement in the task of answer sentence selection and a good achievement in the task of paraphrase determination.

Xiao Li & Qingsheng Li (2015) proved that by applying algorithm based on syntaxes structure and performing semantic relations measurement on the syntaxes structure has succeeded in improving the effectiveness level of sentence similarity measures. More interestingly, only by modifying and combining several semantic relations measurement methods, it is able to increase result of similarity accuracy (Ptáček, 2012).

Meanwhile, hybrid approach combines semantic, corpus, ontology and relations-based approaches (Sumathy & Chidambaram, 2016). Li et al. (2006) showed a measurement technique for sentence similarity based on semantic information and words order information that exist in a sentence. Firstly, semantic similarity is acquired from comparison between raw semantic vector and semantic vector with lexical and corpus database. Subsequently, word order similarity is produced through the comparison of both the vector sets. Finally, sentence similarity is measured by combining semantic similarity and word order similarity.

Hybrid approach is frequently utilized in assessing short essay sentence similarity (Sumathy & Chidambaram, 2016; Li et al., 2006; Pawar & Mango, 2018). Indeed, the latest research in measuring sentence similarity still applies hybrid approach. Pawar & Mango (2018) calculated similarity among words based on side-based approach. Information content from Wordnet lexical database is believed to be able to influence similarity in a specific domain. Semantic vector that contains similarity between words will form sentence and used for sentence similarity calculation. Word order vector will be constructed as well, to calculate the impact of syntaxes structure on a sentence. Sentence similarity is calculated based on both the semantic vector and word order vector.

In a research that involves semantic similarity measurement for phrase and short sentence translation from Arabic to English, Machine and Dictionary Translation techniques has been used (Salha Alzahrani, 2016). Maximum-Translation Average Algorithm applies phrase set produced from Dictionary-Based Technique. Phrase vector and N-V attained from Machine Translation Technique is used to calculate semantic similarity. However, due to sentence similarity measurement only involved semantic equivalence measurement without considering to sentence structure, the issue of word ambiguity exists.

Kadupitiya et al. (2016) used semantic similarity measurement technique (corpus- and knowledge-based similarity measurement). The technique applies semantic relation similarity concept just as Wordnet taxonomy, apart from using word order information. Nevertheless, research shows the result obtained can be optimised further if word ambiguity issue can be overcome by considering surrounding word to acquire sentence context information.

The latest researches on other languages also involve determination of statistical and semantic features in sentence similarity measurement in Portuguese. Four fundamental features used are TF-IDF, Word2Vector, Binary Matrix Method and sentence length (Anderson Pinheiro et al., 2017). TF-IDF is a statistical method to measure significance level of a word in a sentence (Salton & Yang, 1973). Word2Vector is an unsupervised model to generate vector representation for each word in word set that intends to measure semantic similarity between words (Rumelhart et al., 1986). Binary Matrix Method uses Matrix-Based Method to calculate similarity among sentences that are determined based on similarity among words (Ferreira et. al 2016). The final features that was applied by Zhao et al. (2014) and Bjerva et al. (2014) is sentence length. It is measured based on total number of words in the shortest sentence divided by total number of words in the longest sentence. For this method, stop words are removed first. Nonetheless, the usage of statistical method and basic features, such as sentence length could not measure similarity from the aspect of context.

Wang et al. (2016) with their latest research, have measured Thai language sentence similarity based on the frequently used method on English, that are Syntaxes Structure and Semantic Vector. Sentence similarity measurement using POS and POS dependability to calculate syntaxes structure similarity, then measures semantic similarity using Word2Vector. The research also shows the result attained a more precise similarity because the measurement not only assessing the aspect of sentence semantic, but also sentence structure information. However, further study is needed by considering sentence structure using linguistic method that is expected to be able to extract role played by argument in the sentence.

Nevertheless, sentence similarity measurement based on sentence structure or word order has not been able to solve word ambiguity problem. Word ambiguity problem could be solved if surrounding words are considered to acquire parts of context information (Li et al. 2006).

Statistical method could not always identify perfect matching without a clear relation or concept between two natural sentences. Numerous approaches have been used to overcome this problem by determining words arrangement and semantic vector assessment, however those approaches could hardly compare sentences that have complex syntaxes structure constructed based on the usage of long words and sentences using various grammar pattern (Lee et al., 2014).

While Lee et al. (2012) & Mandreoli (2002) applies semantic method. The method applies semantic network, such as Wordnet, Vector Space Model and Statistical Corpus to calculate semantic similarity between words using different measurement methods. Nonetheless, semantic method measures sentence similarity only based on semantic similarity between words, where other syntaxes information and syntaxes knowledge, such as semantic class and thematic role are ignored (Wafa Wali et al., 2017).

To overcome this problem, researchers suggest hybrid method to calculate sentence similarity by considering both semantic information and syntaxes information. Nevertheless, the hybrid method has several weaknesses, such as semantic measurement is made in separation where semantic similarity is calculated based on words semantic similarity, while phrase matching, word order and words occurrence frequency are calculated for syntaxes similarity. Indeed, some knowledge features are not considered in sentences similarity measurement, such as thematic role, semantic class and relation between levels of syntaxes and semantics based on semantic predicate (Wafa Wali et al., 2017). When two sentences that have similar syntaxes structure (subject + verb + object) but semantic class for those arguments are different, the sentence pair is syntactically similar based on hybrid method, whereas according to human expert these two sentences are actually different. For example, the sentences ‘Ali reads a book’ and ‘Ali has a book’ have similar syntax structure and semantic relations for each argument (subject (noun) + verb + object (noun)) exists in both sentences, but different in thematic role played by verb, consequently caused both sentences in actual are not similar at all.

Lee et al. (2014) have applied corpus-based ontology to calculate similarity relations between two words and grammar rule in striving to identify sentence context. Meanwhile, Wafa Wali et al. (2017) have used semantic knowledge technique to determine semantic similarity and syntactico-semantic knowledge technique to reduce the problem of words ambiguity. Both methods are linguistic method that applied grammar rule in managing words ambiguity issue by ascertaining the context of sentence.

Thematic Role

Theta or thematic role refers to the semantic relationship between verb and his argument (Ramli, 2006). For example, verb *menghurai* (parse) requires two arguments marked by its thematic role. The subject argument is labelled as the Agent while the object argument is labelled as the Experiencer.

1. Type of verb

Ramli (2006) has assigned seven significant thematic roles influenced by the present of particular verbs namely Agent, Sufferer, Themes, Location, Beneficiary, Experiencer and Source. Sri Liyaningsih & Siti Zuhriah (2016) stated that there are 12 thematic roles in their arguments (addition of Tool, Number, Purpose, Reason and Time). However, the number and types of thematic roles selected are based on the training and test data sets in this study. All of these thematic roles can be described as follows:

(i) *Penganalisis sintaksis menjelajah token.*
Syntax analyst explore tokens.
AGENT V SUFFERER
(ii) *Pengkompil merupakan penterjemah.*
Compiler is a translator.
THEME V THEME
(iii)*Nahu bebas konteks mewakili peraturan sintaks sesuatu bahasa.*
Context-free-grammar represents the syntax rules of a language.
THEME V EXPERIENCER

Thematic role labelling is considered to occur at the basic structure level. This is mean that the role of specific argument remains same when active sentences are changed to passive sentences. For instance, noun *Penganalisis sintaksis* (Syntax analyst) in sentence (i) will remain as the Agent even if the sentence is converted to the *Token dijelajah oleh penganalisis sintaksis* (Token explored by the syntax analyst).

2. Morphology

Thematic role labelling is determined by predicate or verb to its argument whether the external argument is the noun subject or the inner argument is the noun object (direct or indirect) (Chomsky 1981). Sentence (i) to (iii) are examples of how the thematic roles are labelled by type of verb that are whether transitive, dual-transitive or non-transitive. However, in Malay, apart from the type of verb factor, the presence of prefix and suffix (morphology) also influences the thematic role of significance arguments. For examples,

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

(ii) *lahu* (idle)
a. *Pemproses sedang melahu* (The processor is idle).
b. *Aturcara melahukan pengkompil itu* (The program idling the compiler).
c. *Aturcara melahukan pengkompil itu di peringkat proses* (The program idling the compiler at processing phase).

Those examples show the different structure of the arguments for the verb *hurai* and *lahu*. The difference is influenced by the element of initial prefixes and suffixes imposed on the verb.

(i)	*hurai* (parse)				
a.	*menghurai* :	V;	1 N	2 N	
b.	*menghuraikan*:	V;	1 N	2 N	3 N
(ii)	*lahu* (idle)				
a.	*melahu* :	V;	1 N		
b.	*melahukan*:	V;	1 N	2 N	
c.	*melahukan*:	V;	1 N	2 N	(3) N

Verb *hurai* is a transitive verb while verb *lahu* is a non-transitive verb. However, the structure of this argument is changed influenced by the addition of morphological element to the particular verb. Indirectly, this morphological element has a significant impact on the thematic role played by the verb in the sentrence.

(i) hurai

a. *Penghurai menghurai ayat*.
Parser parse the sentence.
AGENT SUFFERER

b. *Penghurai menghurai ayat kepada token*.
Parser parse the sentence into tokens.
AGENT SUFFERER BENEFICIARY

(ii) lahu

a. *Pemproses sedang melahu*.
The processor is idle.
EXPERIENCER

b. *Aturcara melahukan pengkompil itu*.
The program idling the compiler.
AGENT SUFFERER

c. *Aturcara melahukan pengkompil itu di peringkat proses.*
The program idling the compiler at processing phase.
AGENT SUFFERER LOCATION

From this discussion, it is clear that thematic role for Malay language is not only influenced by the type of verb itself (whether transitive or non-transitive), but the addition of prefix and suffix also has a significance impact on the sentence argument structure and affects rule of Thematic Role. Hence, in this study, rules of Thematic Roles preliminary developed based on these two main factors and at the same, will determine any other factor which will reflect role of subject and object’s arguments.

Semantic Network

Word or phrase match is made based on semantic network information in the Wordnet lexical database by mapping between Malay and English *synsets*.

1. Lexical Database

Lexical is related to the word or vocabulary of a language. A lexical unit is a single word, part of a word or chain of words that forms the basic element of a lexicon of a language, called vocabulary. Lexical databases store lexical information for each word. Lexical information consists of lexical categories and synonyms, including semantic and phonological relations between words or sets of words. For this study, the lexical database included the lexical categories of Nouns, Verbs, Adjectives, and Prepositions.

Lexical databases are different from dictionaries. It takes into account the meaning of the word, whereas the dictionary only considers the list of words. In other words, the lexical database relies on understanding the context of the sentence by extracting the semantic relationship of each significant word in a sentence.

2. Semantic Network and Semantic Relationship

Semantics with a simple definition are meanings. But to understand the actual meaning of a sentence, it is necessary to delve into the actual meaning of each word at first. The actual meaning of each word is derived by mapping each other's semantic relations and processing them until they are understood.

The semantic relationship of each word in a sentence is measured and processed using a semantic network. The semantic network collects words into a set of synonyms known as *synsets*. It also provides a brief definition, example of usage and records the number of relationships between sets of synonyms or their members. Verb, Noun, Adjective and Preposition will be grouped into sets of cognitive synonyms, each one of it presenting a different concept. *Synsets* will be intertwined between semantic concepts and lexical relations. The result of this semantic network is used as one of the methods of measuring sentence similarity by making comparisons between sentences in terms of concepts, not in terms of direct meaning.

3. Semantic Network Measurement Methods

The semantic network similarity measurement uses the information found in the concept of hierarchy (or *synset*) is-a, and calculates the rate of similarity between concept A and concept B (Ted Petersen et al, 2004). For example, measurements will show that apples are more similar to grapes than cars, based on the fact that apples and grapes share fruit as ancestors in the noun hierarchy. However, Noor Syakirah Ibrahim et. Al. (2011) suggest that semantic relations in Wordnet architecture are not only limited to 'is-a' to 'synonyms', but also to semantic relations ‘antonym’, ‘hyponym’ dan ‘hypernym’, ‘meronym’, ‘holonym’ and ‘troponym’. However, is-a relationship is the most widely used semantic relationship in Wordnet (Thabet Slimani, 2013).

Pola Grammar

The term pola refers to the 'variant' which is the 'variant of the sentence' (Mohd Juzaiddin et al., 2006). Pola grammar is a technique to extract features of syntactic and grammatical relationship from structure of the Malay language sentence (Mohd Juzaiddin, 2008). Construction of a sentence is determined by the position of the Malay language subject and predicate’s argument. According to Asmah (2009), there are seven Malay language grammar patterns have been outlined (Juzaiddin Mohd et al., 2006):

(i) Actor + Verb
(ii) Actor + Verb + Complement
(iii) Verb + Complement
(iv) Signified + Signified
(v) Classified + Classifier
(vi) Complement + Verb + Actor
(vii) Complement + Verb

This study has benchmarked the sentence similarity results in short essay assessment based on Pola Grammar Technique for comparison and baseline purpose due to two main factors:

(i) Pola Grammar Technique is a linguistic short essay assessment technique. This technique identifies the position of the subject, verb and object in short sentences, compound sentences and complex sentences. Next, each of these arguments is matched to the arguments contained in the answer scheme. A similar approach is applied using rules of Thematic Role. However, in contrast, these rules make a contextual comparison of the arguments in the sentence.
(ii) Thematic Role rules in this study use the same data set used by Mohd Juzaiddin (2008) which applied Pola Grammar Technique. In order to produce reliable results comparisons, one training data set and three sets of the same test data were used and a fair result comparison was generated.

RESEARCH DEVELOPMENT METHODOLOGY

This study measures Malay sentence similarity through the combination of linguistic method: Thematic Role Technique, and statistical method, Semantic Network Technique. Both techniques are the main process after input, consists of question documents, answer scheme documents students' answer documents are entered and as a final result, similarity marks between students' answer and answer scheme is obtained.

Based on Figure 1, the overall measurement process of sentence similarity between students' answer and answer scheme are divided into three phases: Input Phase, Process Phase and Output Phase.

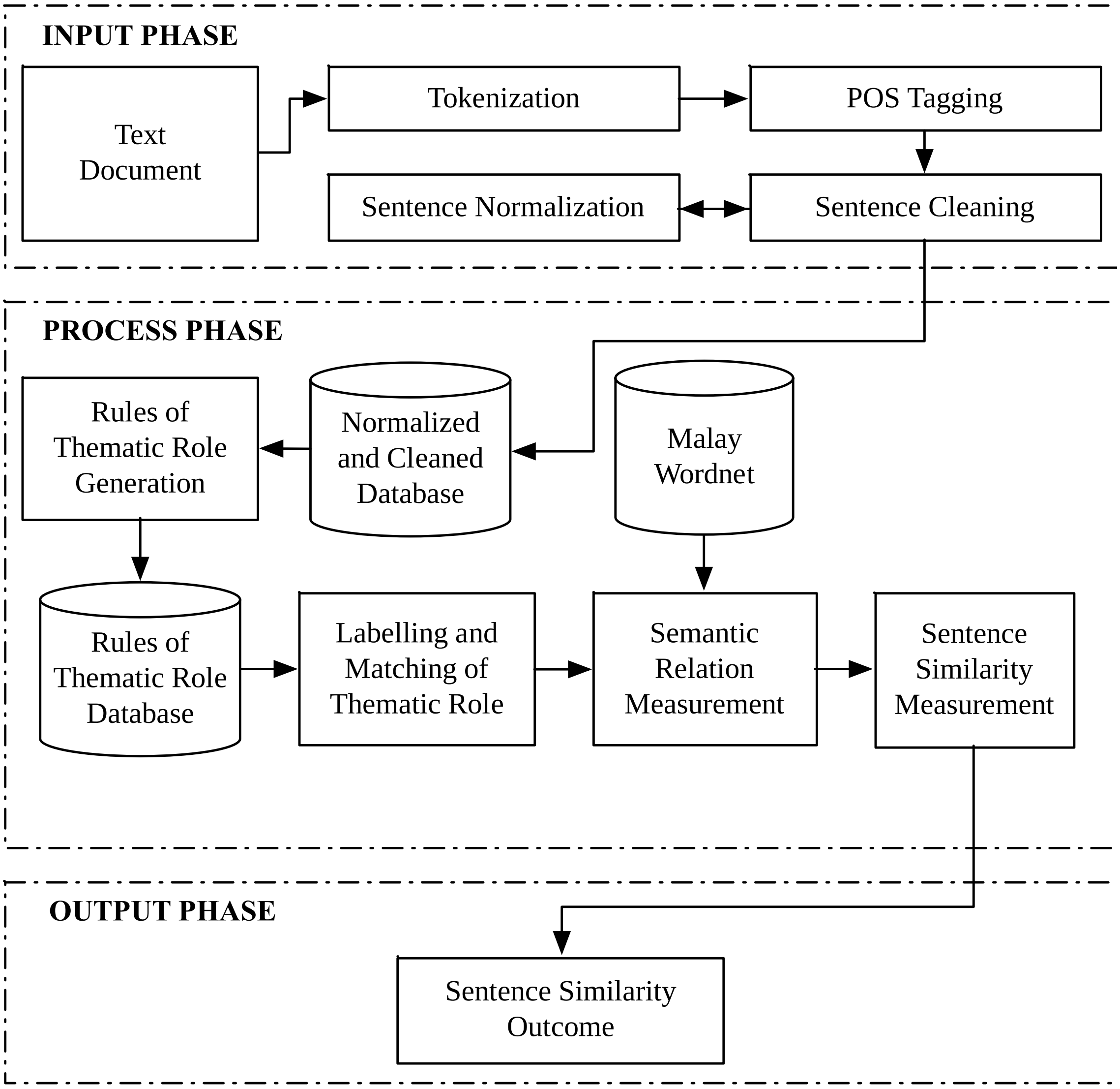

INPUT PHASE
Text Document
Tokenization
POS Tagging
Sentence Normalization
Sentence Cleaning
PROCESS PHASE
Rules of Thematic Role Generation
Normalized and Cleaned Database
Malay Wordnet
Rules of Thematic Role Database
Labelling and Matching of Thematic Role
Semantic Relation Measurement
Sentence Similarity Measurement
OUTPUT PHASE
Sentence Similarity Outcome

Figure 1. The architecture of the measurement process for sentence similarity.

Input Phase

Most similarity measurement processes are only determined based on answer scheme and students' answer question documents, however in this research, question document is considered as well. This is because target answer for certain question, sometimes does not need complete sentence that has both subject and predicate, but only has one of the parts.

1. Tokenization

At the early stage of Input Phase, all three types of documents will go through tokenizing process. Tokenizing is a process, where a sentence will be chunked into smaller parts consist of words and symbols. Tokenizing is preliminary applied on question document. In this process, interrogative words (question words) and type of word after that will be identified to ascertain target answer part.

Subsequently, tokenizing is employed on both answer scheme and students' answer documents. After all sentences have been fragmented into token, it will be merged to construct certain phrases based on two rules; proper noun and compound word.

2. POS Tagging

POS tagger will tag all particular tokens. Once each token is tagged, compound phrase that are exists in the sentence will be identified. In this study, POS tagging are based on those Malay POS tagger (Mi-POS) (Xian et. Al., 2016). This tagger was developed based on a machine learning approach called MaxEnt Model that uses OpenNLP open code NLP equipment. A total of 28 types of words are marked using this tagger. However, there are four types of lexical found to have high implications in IEA; nouns, verbs, adjectives and adjectives.

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. Apart from that, there are compound proper nouns already exist in Malay lexical database.

3. Sentence Normalization

After tagging on words and phrases are complete, 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' with affix insertion meN + verb and discarding the word 'ialah' , 'adalah' and 'merupakan' if there are a more significant verb presence in a sentence.

4. Sentence Cleaning

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. If this happens, it may affect the final result of this study to measure sentences similarity between students' answer and answer scheme.

Process Phase

Process Phase is the most important part in measuring the sentence similarity. This phase involves two main elements: training and testing. For training purposes, Training Data Set A which consists of 10 sets of questions (10 sentences), 10 sets of answer scheme (11 sentences) and 10 sets of answer (71 sentences) comprising all types of simple and compound sentences used to produce rules of thematic role and optimum threshold value for *synset* (token) and sentence similarity. Meanwhile, for testing purposes, Testing Data Set B, C and D that consists of 3 questions sets (3 sentences) and 3 answer scheme sets (9 sentences) and 185 students' answer sets (354 sentences) that comprise of various simple and compound types of sentences have been utilized.

1. The Generation of Rules of Thematic Role

Rules of Thematic Role is generated by conducting training on Training Data Set A that has question, answer scheme and students' answer documents acquired from Database of Normalized and Cleaned Data Set. Most of similarity measurement processes are only ascertained by answer scheme set and students' answer documents, however in this study, question document is considered as well. This is because target answer for certain questions sometimes does not require a complete sentence that has both arguments: subject and predicate, but only needs one of the arguments. For example:

Question: *Apakah tugas penghurai? (What is the task of parser?)*
Complete answer: *Pengkompil menghurai token. (Parser parse the token.)*
[subject] [predicate]
Target answer: *Menghurai token. (Parse the token.)*
[predicate]
Non-target answer: *Penghurai. (Parser.)*
[subject]

Based on the example, the question: '*Apakah tugas penghurai?*' aims the answer towards predicate, rather than subject. In other words, if the student's answer only contains predicate (which is similar), thus it will be measured as similar even though the sentence is incomplete because it has no subject. Based on the question, it is found that type of question words and token after question word influence target answer, it exists in either subject, verb or object, or combination of one these. This assists significantly when answer given by students only answer directly with verb and object, whereas it did not fulfil requirement for valid sentence structure. However, if the answer for the question is contained in the predicate part (verb + object), thus the answer should be accepted as correct although without the presence of subject.

The result of training has generated eight significant thematic roles. Rules of thematic role which were previously produced is used on answer scheme document to label and identify points that exist in each answer scheme. Further, it will be applied on students' answer data set to test the effectiveness of this method against human assessment. The generated rules of thematic role will be saved in Rules of Thematic Role Database (as shown in Table 1).

Table 1. Rules of Thematic Role.

Morphology	Special Word	Rules of Thematic Rols
meN..., meN-...-kan, memper-...-i, memper-...-kan	-	PEL-KK-PEND
	kepada	PEL-KK-PEND-(kepada)TEMA/PEM
	iaitu	PEL-KK-PEND-(iaitu)TEMA
	daripada, kepada	PEL-KK-PEND-(daripada)SUMBER-(kepada)TEMA/PEM
	dari-ke	PEL-KK-PEND-(dari)SUMBER-(ke)TEMA/PEM
	kepada, dalam	PEL-KK-PEND-(kepada)TEMA/PEM-(dalam)TMPT
	kepada	PEL-KK-PEND-(kepada)TEMA/PEM
	dalam	PEL-KK-PEND-(dalam)TMPT
	pada	PEL-KK-PEND-(pada)TMPT
	semasa	PEL-KK-PEND-(semasa)MASA
me-…-i	-	PEL-KK-PENG
di-…	oleh	PEND-KK-PEL
	ke/kepada	PEND-KK-(ke)TEMA/PEM
	semasa	PEND-KK-(semasa)MASA
	oleh, dalam bentuk	PEND-KK-PEL-(dalam bentuk)TEMA
di-…kan	mengikut/berdasarkan	PEND-KK-(mengikut)TEMA
	semasa	PEND-KK-(semasa)MASA
	daripada	PEND-KK-(daripada)SUMBER
	untuk	PEND-KK-(untuk)PEM
	sebagai	PEND-KK-(sebagai)TEMA
	dalam	PEND-KK-(dalam)TMPT
menerima, mendapat, memperoleh	dalam bentuk	PEM-KK-TEMA-(dalam bentuk)TEMA
menjadi	-	PENG-KK-TEMA
mempunyai, terdapat	-	PEM-KK-TEMA
	dalam	PEM-KK-TEMA-(dalam)TMPT
iaitu, bertindak sebagai	-	TEMA-KK-TEMA
	kepada, dalam	TEMA-KK-TEMA-(kepada)PEM/TEMA-(dalam)TMPT
merupakan, adalah, ialah, berupa, terdiri daripada	-	TEMA-KK-TEMA
	kata bilangan	TEMA-KK-TEMA-(BIL)
	dalam	TEMA-(dalam)TMPT-KK-TEMA
	bagi	TEMA-KK-TEMA-(bagi)PEM
digunakan	untuk	PEND-KK-(untuk)PEM
	dalam	TEMA-KK-(dalam)TMPT
	pada	TEMA-KK-(pada)TMPT
	oleh	PEND-KK-(oleh)PEL
mengikut	-	PEND-KK-TEMA

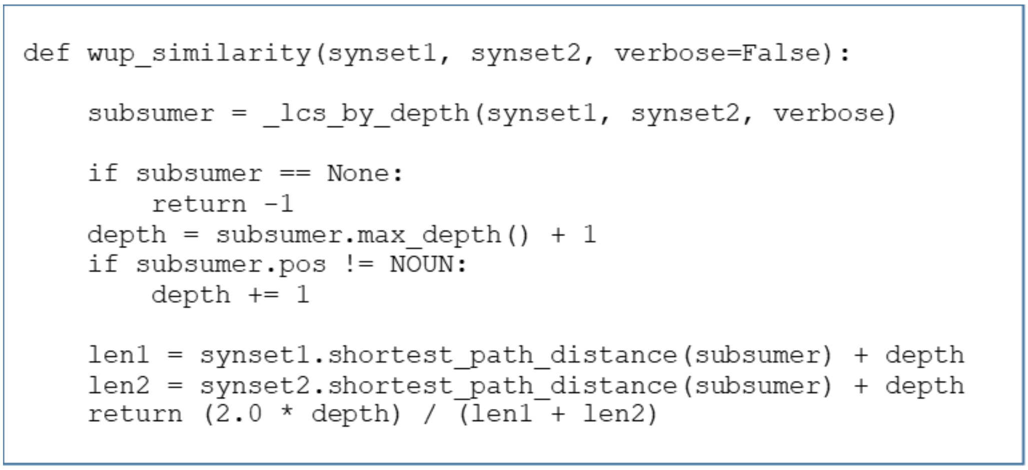
2. Labelling and Matching of Thematic Role

Labelling of thematic role is applied on students' answer document by referring to the Rules of Thematic Role Database. For simple sentence, labelling of thematic role is quite easy because thematic role of an argument does not change. Contrarily for compound sentence, thematic role for subject argument for example, can change according to several verb contained in the sentence construction.

After labelling of thematic role is conducted, matching process of rules pattern on students' answer document is implemented by matching the rules of thematic role that exists in the students' answer document set against answer scheme document. If match exists, it is assumed as matching. However, if more than one matches for similar thematic role pattern against answer scheme set, all the matches will be counted for subsequent similarity measurement, which is semantic relation measurement.

3. Semantic Relation Measurement

The semantic relation measurement is performed on words and phrases labelled with thematic role. The measurement is conducted using modified Wu & Palmer's (wup) method that calculates relativity rate by considering depth of two *synsets* based on LCS in Malay Wordnet taxonomy (map to English wordnet). Modified wup is chose to use in this research by considering of its feature which is less complexity to implement in a pervasive computing system where the context is modelled using an ontology and gives realistic similarity results (Guessoum, 2016).



```
def wup_similarity(synset1, synset2, verbose=False):

    subsumer = _lcs_by_depth(synset1, synset2, verbose)

    if subsumer == None:
        return -1
    depth = subsumer.max_depth() + 1
    if subsumer.pos != NOUN:
        depth += 1

    len1 = synset1.shortest_path_distance(subsumer) + depth
    len2 = synset2.shortest_path_distance(subsumer) + depth
    return (2.0 * depth) / (len1 + len2)
```

Figure 2. The algorithm of semantic similarity measurement based on LCS method.

In *wup_similarity()* shown in the algorithm in Figure 2, it will return a value that portrays semantic relation similarity level between two *synsets* (words or phrases), based on LCS method (the most specific ancestor node). LCS not necessarily to be calculated based on the shortest path that connects two semantic relations, it conversely considers the deepest common ancestor based on taxonomy. If there are several LCS choices, the longest path with root not will be chosen. The longest path will be selected for calculation.

Decimal value produced from similarity measurement between $synset_1$ and $synset_2$ is in the range $0 \geq similarity\ value \geq 1$. If the path that connects between the two semantic relations is not found, thus value -1 will be returned and the longest path from LCS to root node is obtained by adding value 1 because the calculation considers both the start nod and end node. Subsequently, the shortest path is acquired from LCS to each *synset* that is sub-added. The resultant is added on the LCS path length to attain path length for each *synset* to root node.

```
def _lcs_by_depth(synset1, synset2, verbose=False):
    subsumer = None
    max_min_path_length = -1

    subsumers = common_hypernyms(synset1, synset2)

    if verbose:
        print "> Subsumers1:", subsumers

    eliminated = set()
    for s1 in subsumers:
        for s2 in subsumers:
            if s2 in s1.closure(HYPERNYM):
                eliminated.add(s2)
    if verbose:
        print "> Eliminated:", eliminated

    subsumers = [s for s in subsumers if s not in eliminated]

    if verbose:
        print "> Subsumers2:", subsumers

    for candidate in subsumers:

        paths_to_root = candidate.hypernym_paths()
        min_path_length = -1

        for path in paths_to_root:
            if min_path_length < 0 or len(path) < min_path_length:
                min_path_length = len(path)

        if min_path_length > max_min_path_length:
            max_min_path_length = min_path_length
            subsumer = candidate

    if verbose:
        print "> LCS Subsumer by depth:", subsumer


    return subsumer
```

Figure 3. The algorithm of LCS calculation based on *synset* depth.

Figure 3 depicts the algorithm for LCS calculation based on depth, *lcs_by_depth()*. This function aims to seek LCS for two *synsets* in Wordnet taxonomy. LCS is defined as common ancestor node for both *synsets* where the shortest path to root node is the longest. *Eliminated* acts to eliminate *synset* that becomes the ancestor to other *synset* in the sub-add set. Finally, the function will calculate the length of the shortest path to the root node for every sub-add set. The longest sub-add is selected. Besides that, for word-to-word match, matching is performed directly. However, for phrase matching, it is done based on the Equation (1) to (3):

Head matches without modifier: $0.8x$ (1)

Head matches with modifier: $0.5\mathrm{x} + 0.5\sum y_n$ (2)

Head does not match with modifier: $0.8\mathrm{x} + 0.2\sum y_n$ (3)

where,

x is content value
y is modifier value

For case (i), content between *synset*$_1$ and *synset*$_2$ are matches. For example, *synset*$_1$ is 'penganalisa' (analyzer) and *synset*$_2$ is 'penganalisa leksikal' (lexical analyzer). Match value is $0.8 \times 1 = 0.8$. For case (ii), *synset*$_1$ is 'penganalisa leksikal' (lexical analyzer) and *synset*$_2$ is 'penganalisa semantik' (lexical semantic). The content of both heads are matches that is 'penganalisa' (analyzer), while the modifier is not match between 'leksikal' (lexical) and 'semantik' (semantic). Therefore, the relativity result is $(0.5 \times 1) + (0.5 \times 0) = 0.5$. For case (iii), head does not match the modifier: *synset*$_1$ is 'penganalisa leksikal' (lexical analyzer) and *synset*$_2$ is 'penterjemah semantik' (semantic interpreter). The relativity result is $(0.8 \times 0.5) + (0.2 \times 0) = 0.4$.

However, this semantic relation similarity measurement based on Wordnet taxonomy is subjected to threshold value. This value is determined during training process on particular data set in the research.

4. Sentence Similarity Measurement

At the end of process, sentence similarity measurement was performed between answer scheme and students' answer without ignoring the sentence structure. Similarity measurement is calculated based on Equation (4):

$$\frac{\sum_{i=1}^{n} S(s_i)}{N} \times w_i \times -w_{neg} \tag{4}$$

where,

n is number of *synset* comparison
S is similarity measurement
s is relativity value of word and phrase
N is number of *synset* in argument structure
w is existence value of word or phrase
neg is negative value of word or phrase.

The equation will take into account types and number of arguments that have been labelled. Only matches argument structure will be processed. In calculating a sentence similarity in students' answer document against sentence in answer scheme document, the total number of *synset* similarity is not the actual measurement. Thus, the existence of every *synset* is considered important because even though a *synset* exists in a sentence, if its similarity is very low (under the *synset* threshold value), the *synset* should not be accepted as similar in the sentence.

However, just as semantic relation similarity measurement of *synset* is subjected to the threshold value of *synset*, sentence similarity measurement also depends on the identified threshold value.

5. Determination of Optimum Threshold Value

There are two parts of similarity where threshold value is set. Both these threshold values are established from the result of training process on training data set. The first part is *synset* similarity measurement part. The optimum threshold value identified was 0.6. On the second part, sentence similarity measurement, 0.75 was the determined optimum threshold value.

Output Phase

The phase has two outcomes: training similarity outcome and test similarity outcome. In training similarity outcome, *synset* similarity threshold value and sentence similarity threshold value was calibrated until the optimum mark of similarity outcome against human valuation was obtained. Whereas, test similarity outcome is where both threshold values were used to evaluate the final outcome of evaluation similarity performed by thematic role and semantic network method against human assessment.

1. Sentence Similarity Outcome

As the research final outcome, the similarity between both students' answer document and answer scheme document were measured. The document consists of one or more sentences. Some answer scheme and students' answer documents only have one argument structure and other has many argument structures. Document similarity was calculated based on number of points and marks for a question. In Table 1, for Data Set A: Question 1, 2, 3, 4 and 6, 1 point is equal to 1 mark. For other question in Data Set A: Question 5 (1 point is equal to 0.2 marks), Question 7 and 8 (1 point is equal to 0.33 marks), Question 9 and 10 (1 point is equal to 0.25 marks), Data Set B (1 point is equal to 0.25 marks), Data Set C (1 point is equal to 0.3 marks) and Data Set D (1 point is equal to 0.5 marks).

The value of the mark obtained was calculated using equation (1.4). For each argument structure that matches and has similarity value, S, and presence of valid, w, the value of mark was filtered based on the threshold value. For argument structure, which is similar or higher than the threshold value, 1 point is given. Finally, the number of points is sum up, n, and final mark value is counted based on the equations in Table 2.

Table 2. The equations for mark calculation of answer set.

Data Set	Equation
Data Set A – Question 1	n
Data Set A – Question 2	n
Data Set A – Question 3	n
Data Set A – Question 4	n
Data Set A – Question 5	$\frac{n}{5}$
Data Set A– Question 6	n
Data Set A– Question 7	$\frac{n}{3}$
Data Set A – Question 8	$\frac{n}{3}$
Data Set A – Question 9	$\frac{n}{4}$
Data Set A – Question 10	$\frac{n}{4}$
Data Set B	$\frac{n}{10} \times 2.5$
Data Set C	$\frac{n}{5} \times 1.5$
Data Set D	$\frac{n}{4} \times 2$

Testing Dataset B

Table 3 to 5 only display some of the outcomes that require evaluation error of more than 0.5 against human assessment using Thematic Role and Semantic Network Techniques and Pola Grammar Technique.

For Testing Data Set B in Table 3, question '*Nyatakan tugas utama suatu pengkompil'* (State the main task of a compiler) is cleaned to '*Nyatakan tugas pengkompil'* (State task of compiler). Next, it is defined to question qord *'Nyatakan'* (State) followed by verb *'tugas'* (task) will aim at predicate which consists of verb and object (noun phrase) instead of subject.

Table 3. The Similarity Result of Data Set B.

Question Number	Human	Pola Grammar (PG)	Thematic Role + Semantic Network (PT+RS)	Assessment Error	
				PG	PT+RS
1	1.5	0.5	1.5	1	0
2	0.2	1.05	0.5	-0.85	-0.3
6	1	0	0.25	1	0.75
8	1	1.65	1	-0.65	0
9	0	0.965	0.25	-0.965	-0.25
10	2	1.17	1.25	0.83	0.75
17	2.5	2.115	1.25	0.385	1.25
19	0.2	0.965	0.25	-0.765	-0.05
26	0	0	1	0	-1
28	0.2	1.76	0.5	-1.56	-0.3
31	0	1.05	1	-1.05	-1
34	1	1.525	1	-0.525	0
37	2	0	1	2	1

As a conclusion to similarity measurement outcome on the data set, precision, recall and f-measure values for all 41 students' answer set on average were as high as 91.3%, 84% and 87.5% using combination of Thematic Role and Semantic Network Techniques compared to 79.17%, 76% and 77.55% using Pola Grammar Technique. Thus, the achievement of using the combination of Thematic Role and Semantic Network Techniques in measuring most simple sentence show better result, which is 10.05% higher on average.

Testing Dataset C

For Testing Data Set C in Table 4, the question '*Berikan definisi ringkas mengenai Nahu Bebas Konteks*' (*Give brief definition of Context Free Grammar)* is cleaned to '*Berikan definisi Nahu Bebas Konteks*' (*Give Context Free Grammar definition).* Subsequently, it will be interpreted to question word 'Berikan' (Give) followed by noun 'definisi' (definition) that will aim at verb and object of noun phrase type instead of subject.

Table 4. The Similarity Result Data Set C.

Question Number	Human	Pola Grammar (PG)	Thematic Role + Semantic Network (PT+RS)	Assessment Error	
				PG	PT+RS
5	1.5	0.944	0.6	0.556	0.9
30	1	0.78	0.3	0.22	0.7
31	0	0.944	0	0.944	0

As a conclusion to the outcome of sentence similarity measurement on this data set, precision, recall and f-measure values for all 34 students' answer set on average were as high as 100%, 90% and 94.74% using combination of Thematic Role and Semantic Network Techniques compared to 95% for all three methods using Pola Grammar Technique. Therefore, the achievement of using the combination Thematic Role and Semantic Network Techniques in measuring most compound sentences in answer set is almost similar to Pola Grammar Technique's performance.

Testing Dataset D

For Testing Data Set D in Table 5, the question '*Berikan definisi token dalam pengkompil*' (*Give token definition in compiler)* will be interpreted to question word 'Berikan'(Give) followed by noun 'definisi' (definition) will aim at verb and object of noun phrase type instead of subject.

Table 5. The Similarity Result Data Set D.

Question Number	Human	Pola Grammar (PG)	Thematic Role + Semantic Network (PT+RS)	Assessment Error	
				PG	PT+RS
2	1	0.412	1	0.588	0
4	2	0.576	1.5	1.424	0.5
5	0.5	1.36	0.5	-0.86	0
14	0.5	1.358	0.5	-0.858	0
16	1	1.812	1	-0.812	0
17	0	1.132	0.5	-1.132	-0.5
18	1	1.586	1	-0.586	0
20	0	0.802	0.5	-0.802	-0.5
22	0	0.802	0	-0.802	0
25	0	0.576	0.5	-0.576	-0.5
26	0	0.576	0.5	-0.576	-0.5
30	0	0.802	0	-0.802	0
34	1	0	0	1	1

There are 43 answer sets in Testing Data Set D. Table 5 displays that all students' answer set were successfully assessed similarly using the combination of Thematic Role and Semantic Network Techniques. In contrast, the Pola Grammar Technique shows 13 answer sets were not successfully assessed similarly against human evaluation. This means similarity percentage of combination Thematic Role and Semantic Network Techniques was 100% against Pola Grammar Technique with 69.77%, a 30.23% advantage.

As a conclusion to the outcome of sentence similarity measurement on this Testing Data Set D, precision, recall and f-measure values for all students' answer set on average were as high as 100%, for all three methods using the combination of Thematic Role and Semantic Network Techniques compared to 76%, 73.08% and 74.51% using the Pola Grammar Technique. Therefore, the achievement of the using both methods in measuring most compound sentences in answer set is excellent compared to Pola Grammar Technique usage.

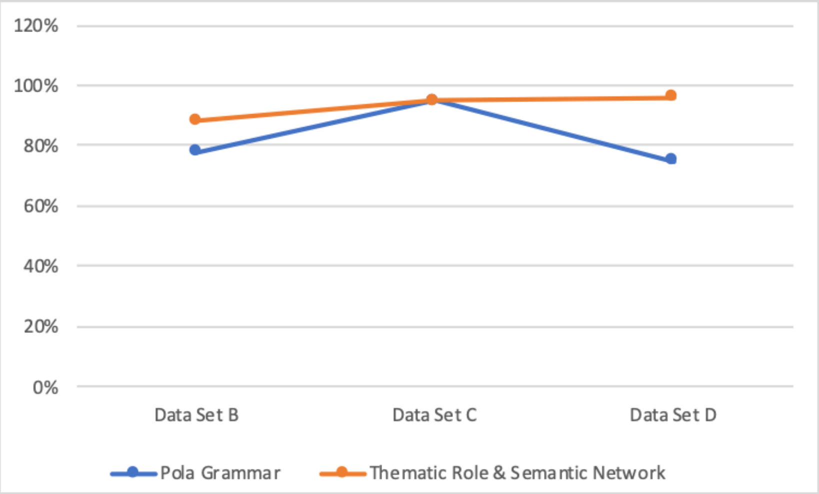


Figure 4. Result of sentence similarity based on *f-measure* method according to data set.

Similarity outcome were performed on three testing data sets: Testing Dataset B, C and D. This is because each data set contains answer set that has particular characteristics of certain sentence types. Testing Data Set B and D consist of fixed compound sentences and mixed compound sentences. Meanwhile, Testing Data Set C has combination of simple sentences and mixed compound sentences. Some of the factors identified which contribute to the mark's dissimilarity for students' answer set using Thematic Role and Semantic Network Techniques are:

i. Semantic relation similarity using Semantic Network Technique considering both super-class and sub-class relations for two *synsets*.
ii. Semantic relations similarity measurement between *synset* in phrase form did not comply with head and modifier rules by human.
iii. Human assessment sometimes does not quite comply with number of points in a sentence in conducting evaluation.
iv. Human assessment sometimes does not consider argument context (thematic role) in a sentence.
v. For incomplete sentence (no subject), some of the answers were not accepted by human assessor and some of the answers were accepted.

Based on the average value of *f-measure* on each data set, in Testing Data Set C, both techniques produced similar assessment accuracy of 95% and 94.74% against human evaluation. Nevertheless, for Testing Data Set B and D, Pola Grammar Technique established assessment accuracy rate of 77.55% and 74.51% compared to human evaluation, a bit low when compared to assessment accuracy using Thematic Role and Semantic Network Techniques with 87.5% and 100% respectively. This means, Pola Grammar Technique is more effective when it is used on answer with simple sentences. For answers in mixed compound sentence, the achievement of both techniques is balanced. However, for evaluation on answers in more complex compound sentences, the usage of Thematic Role and Semantic Network Techniques was proven better.

Table 6. The average of similarity outcome.

Testing measurement method	Pola Grammar (PG)	Thematic Role + Semantic Network (PT+RS)
Precision	83.39%	95.83%
Recall	81.36%	91.33%
f-measure	82.36%	93.53%

As a final conclusion, referring to Table 6, the average of precision, recall and f-measure values on all three test data sets by using Pola Grammar Technique are 83.39%, 81.36% and 82.36% compared to using Thematic Role and Semantic Network Techniques with 95.83%, 91.33% and 93.53%.

CONCLUSION

The study was performed based on the main objective that uses the Rules of Thematic Role that has been developed to verify sentence structure in students' answer set and applies Semantic Network based on Wordnet architecture to measure similarity in two stages: *synset* similarity stage and sentence similarity stage. The constructed rules of thematic role were used to make thematic role labelling for each relevant argument (noun phrase subject, verb and noun phrase object) on answer scheme and students' answer documents at testing stage of the study. The significant number and types of arguments depend on set of rules that was constructed. The thematic role labelling aims to identify sentence structure in question, answer scheme and students' answer documents. With this labelling, the sentence context based on argument thematic role that has been labelled can be determined.

For the purpose of similarity measurement between students' answer and answer scheme, semantic relations between two *synsets* consist of noun phrase, verb phrase, adjective phrase and function phrase, were measured. Wordnet architecture allows semantic similarity and relativity rate measurements between two concepts (word meaning) to be obtained. The architecture also provides six similarity measurements and three types of relativity rates measurements based on Wordnet lexical database. Due to the research attempted to measure context-based sentence similarity, semantic similarity measurement was selected using modified Wu & Palmer's (wup) method.

Wup method seeks path length to root nod based on least common subsummer (LCS) or shortest length for two *synsets* (concept). In other words, the most specific concept was shared as ancestor. The similarity value was acquired by calculating the total of path length from concept to root. The outcome for the similarity measurement on each answer set in all three testing data sets: Data Set B, C and D were compared with the outcome of similarity measurement using Pola Grammar Technique. The precision, recall and f-measure measurement methods were used and as comparison, the usage of Thematic Role and Network Semantic Techniques produced similarity measurement with precision, recall and f-measure averages were found to be 95.83%, 91.33% and 93.53% against 83.39%, 81.36% dan 82.36% for Pola Grammar Technique. This exhibits the average increase for assessment approximation rate was 11.17%. It can be concluded that the combination of Thematic Role and Semantic Network Techniques can perform a better Malay short essay assessment on all types of sentences.

REFERENCES

Anderson Pinheiro, Rafael Ferreira, Máverick André, D. Ferreira, Vitor B. Rolim & João Vitor S. Tenório. (2017). Statistical and Semantic Features to Measure Sentence Similarity in Portuguese. 2017 Brazilian Conference on Intelligent Systems, 342-347.

Asmah Haji Omar. (2009). *Nahu Melayu Mutakhir*. 5th edition. Kuala Lumpur: Penerbit Dewan Bahasa dan Pustaka.

Attali, Y. & Burstien, J. (2006). Automated Essay Scoring with E-rater V.2. *Journal of Technology, Learning and Assessment (JTAA), 4*(3).

Bjerva, J., Bos, J., Goot, R. V. D. & Nissim, M. (2014). The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. 2014 International Workshop on Semantic Evaluation, 642-646.

Burstein, J. & Wolska, M. (2003). Toward Evaluation of Writing Style: Finding Overly Repetitive word use in student essays. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, 35-42.

Chomsky, N. (1981). *Lectures on Government and Binding*. Italy: Foris Publications.

Diana Perez, M. (2004). *Automatic Evaluation of Users' Short Essays by Using Statistical and Shallow NLP Techniques*. Advanced Studies Diploma Work. Universidad Autónoma de Madrid.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment 5*(1).

Erickson, J. D. (2000). *Using Keywords and Computers to Assess Student Writing*, Dissertation of PhD, Washington State University.

Ferreira, R., Lins, R. D., Simske, S.J., Freitas, F. & Riss, M. (2016). Assessing sentence similarity through lexical, syntactic and semantic analysis. *Journal of Computer Speech & Language 39*(C), 1-28.

Fitzgerald, K.R. (1994). Computerized Scoring? A Question of Theory and Practice. *Journal of Basic Writing 13*(2), 3–17.

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

Guessoum, D., & Miraoui, M., & Tadj, C. (2016). A modification of Wu and Palmer Semantic Similarity Measure. *The Tenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*.

Gustafson, N., Pera, M., & Ng, Y. (2008). Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity. Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008, 690-696.

Hearst, M. A. (2000). The debate on automated essay grading. IEEE Intelligent Systems.

Higgins, D., & Burstein, J. (2007). Sentence Similarity Measures for Essay Coherence. Proceedings of the 7th International Workshop on Computational Semantics (IWCS).

Ho, C. Masrah Azrifah Azmi Murad, Shyamala C. Doraisamy & Rabiah Abdul Kadir. (2010a). Measuring Sentence Similarity from Both the Perspectives of Commonalities and Differences. 22nd International Conference on Tools with Artificial Intelligence, 318-322.

Ho, C., Masrah Azrifah Azmi Murad, Rabiah Abdul Kadir & Shyamala C. Doraisamy. (2010b). Word Sense Disambiguation-based Sentence Similarity. 23rd International Conference on Computational Linguistics, 418-426.

Islam, A. & Inkpen, D. 2008. Semantic Text Similarity using Corpus-based Word Similarity and String Similarity. *Journal ACM Transactions on Knowledge Discovery from Data (TKDD) 2*(2), 10.

Kadupitiya, J., Ranathunga, S., & Dias, G. (2016). Sinhala Short Sentence Similarity Measures using Corpus-Based Similarity for Short Answer Grading. Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing, 44-53.

Landauer, T.K., & Laham, D. (2000). The Intelligent Essay Assessor. IEEE Intelligent Systems, 15(5), 27–31.

Leacock, C. & Chodorow, M. (2003). *C-rater: Automated Scoring of Short-Answer Questions*. Language Resources and Evaluation, 389-405, Netherlands: Springer.

Lee, M. C. (2010). A novel sentence similarity measure for semantic-based expert systems. *Journal Expert Systems with Applications: An International Journal archive* 38(5): 6392-6399.

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

Li, Y., McLean, D., Bandar, Z. A., O'shea J. D., & Crockett, K. (2006). Sentence Similarity Based on Semantic Nets and Corpus Statistics, *Journal IEEE Transactions on Knowledge and Data Engineering 18*(8), 1138-1150.

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.

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

Mohd Juzaiddin Ab Aziz, Fatimah Ahmad, Abdul Azim Abd. Ghani & Ramlan Mahmod. (2006). Pola Grammar Technique for Grammatical Relation Extraction in Malay Language. *Malaysian Journal of Computer Science 19*(1), 59-72.

Mohd Juzaiddin Ab Aziz, Fatimah Dato' Ahmad, Abdul Azim Abdul Ghani & Ramlan Mahmod. (2008). Identify Malay Sentence Similarity based on Pola Grammar Algorithm, 12th WSEAS International Conference on Computers, 922-927.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 752-762.

Ning, C., Wang, R., Chen, Z., & Lu, B. 2011. An Efficient Similarity Measure Algorithm of Chinese Sentence. International Conference on Computer Science and Automation Engineering 2011 IEEE (CSAE), 387-390.

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.

Pawar, A., & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. IEEE Transactions on Knowledge and data Engineering.

Ptáček, T. (2012). *Advanced Methods for Sentence Semantic Similarity*. Thesis of Master. University of West Bohemia, Faculty of Applied Sciences Department of Computer Science and Engineering.

Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic Short Answer Marking. Proceedings of Second Workshop on Building Education Using NLP, 9-16.

Raheel Siddiqi. (2010). *Improving Learning and Teaching Through AMS Answer Marking*. Thesis of PhD. Faculty of Engineering and Physical Sciences, The University of Manchester.

Ramalingam, V. V., Pandian, A., Prateek Chetry & Himanshu Nigam. (2018). Automated Essay Grading using Machine Learning Algorithm. *Journal of Physics Conference Series 1000*(1), 012030.

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

Rumelhart, D.E, Hintont, G.E & Williams, R. J. Learning representations by back-propagating errors. (1986). Nature, 323(6088), 533–536.

Salha Alzahrani. (2016). Cross-Language Semantic Similarity of Arabic-English Short Phrases and Sentences. *Journal of Computer Science 12*(1), 1-18.

Salton, G. & Yang, C.S. (1973). On the Specification of Term Values in Automatic Indexing. *Journal of Documentation 29*(4), 351-372.

Shan, J., Liu, Z., & Zhou, W. (2009). Sentence Similarity Measure Based on Events and Content Words. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 623-627.

Shermis, M. D. & Burstein, J. (2016). *Automated Essay Scoring: A cross disciplinary perspective*. Applied Psychological Measurement, United Kingdom: Routledge.

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.

Sumathy, K. L., & Chidambaram. (2016). A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *7*(8), 231-237.

Thabet Slimani. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications 80*(10), 25-33.

Valenti, S., Neri, F. & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education 2*, 319-330.

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.

Wang, J. & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *The Journal of Technology, Learning, and Assessment 6*(2), 1-28.

Wang, Z., Mi, H., & Ittycheriah, A. (2016). Sentence Similarity Learning by Lexical Decomposition and Composition. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, 1340–1349.

Xian, B., Lubani, M., Liew, K., Bouzekri, K., Mahmud, R., & Lukose, D. (2016). Benchmarking Mi-POS: Malay Part-of-Speech Tagger. *International Journal of Knowledge Engineering, 2*, 115-121.

Xiao Li & Qingsheng Li. (2015). Calculation of Sentence Semantic Similarity Based on Syntactic Structure. Mathematical Problems in Engineering Volume 2015, Article ID 203475, 8 pages http://dx.doi.org/10.1155/2015/203475.

Zhao, J., Zhu, T. T., & Lan, M. (2014). ECNU: One Stone Two Birds: Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 271-277.