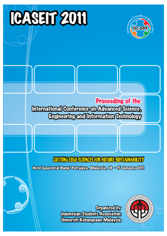
Proceeding of the International Conference on Advanced Science, Engineering and Information Technology 2011

Hotel Equatorial Bangi-Putrajaya, Malaysia, 14 - 15 January 2011

ISBN 978-983-42366-4-9



A Synonym Contextual-based Process for Handling Word Similarity in Malay Sentence

Suhaimi Ab Rahman[#], Nazlia Bt Omar[*], Hassan Mohamed[#], Mohd Juzaidin Ab Aziz [*]

[#] Software Engineering Department, Universiti Tenaga Nasional
Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
Tel.:+603 8921 2020, E-mail: smie@uniten.edu.my;MHassan@uniten.edu.my

**School of Computer Science, Faculty of Information Science and Technology*
Universiti Kebangsaan Malaysia(UKM), Bangi, 43600, Selangor, Malaysia
Tel.:+603 8921 6179, E-mail: no@ftsm.ukm.my;din@ftsm.ukm.my

***Abstract*— In this paper, we attempt to describe a method of finding word similarity within a Malay sentence. The list of similarity word produced is based on searching the appropriate context within a Malay sentence. The context is determined by seeking rules from a rule-based phrase database. In implementing this approach, a working prototype application is described which can be used as a tool for improving writing text in Malay language, especially well adapted toward the requirements of teaching and learning this language in primary and secondary schools. The overall concept presented in this paper will assist us to identify clearly what are the basic components and their specifications that should exist in the process. On the other hand, it is also important to point out the possible drawbacks and constraints of the practical approach suggested.**

***Keywords*— synonyms, rule-based, Malay grammar structure, Malay Part-of-Speech(POS).**

I. Introduction

The meaning of synonym refers to the word that has similar word meaning in the list. As noted in [12], a similarity between two words is often represented by similarity between concepts associated with the two words. A number of semantic similarity methods have been researched and developed in the past, such as in [2,8,10] and different similarity methods have proven to be useful in some specific applications of computational intelligence. Extensive research and review on the existing available methods will help us to identify current trends and issues of relevant approach used and will be a good idea on how we can ratify them into our study. For extensive comparisons of some representative similarity measures, we are referred with [1] and [2]. In general, these methods can be categorized into two groups: edge counting-based (or dictionary/thesaurus-based) methods and information theory-based (or corpus-based) methods. In our research work, we intend to use a corpus of data from a general domain collected from a Kamus Dewan Bahasa dan Pustaka (DBP). We are manually choosing the Malay sentences and words that should be in our data requirement and analysis. The more data collected, the better results can be produced.

Generally, the outcome of this paper is to discuss an overview of the proposed system and and its components, focussing more on the flow of the process. It is important that our processes reflect the functional behaviour of the activities. For this reason, we need to choose wisely how to represent each activity and how these activities will behave.

To further describe, we will organize this paper as follows: Section II shortly discusses on research related with our research work. Section III briefly describes our system flow of the process. Finally, section IV concludes the paper and further research.

II. RELATED RESEARCH

According to [6], the word similarity will measure based on the distibutional pattern of words. The approach used to handle automatic detection of similar words from text corpora is using dependency relationship as the word features, based on which word similarities are computed. The evaluation results are compared against word similarity measures based on WordNet and Roget. For our proposed approach, we will construct a standard form of rule-based pattern of words that can analyse a Malay sentence based on the four main categories of phrases, for instance *"Frasa Nama(FN)"* (Noun Phrase), *"Frasa Kerja(FK)"* (Verb Phrase), *"Frasa Adjektif(FA)"* (Adjective Phrase) and *"Frasa Sendi Nama(FS)"* (Prepositional Phrase).

As discussed in [13], semantic relationships between contextual synonyms can be measure based on the following three semantic relationships, such as Embedment, Intersection and Non-coherence.

As noted in [11], the word clustering process is conducted based on verbs and their objective nouns. While in [7], the broad-coverage parser method extracts dependency triples from the text corpus. A dependency triple consists of two words and a grammatical relationship between them. This structure is important in order to determine the word clustering from the text or sentence.

III. SYSTEM FLOW OF THE PROCESS

Process flow diagrams should include the information regarding the connection between various components or modules. This is to ensure leased capabilities are supported adequately and can achieve specified availability requirements into our study. Fig. 1 shows a diagram on how the components can work with each other in order to complete the series of process in our earlier research planning phase.

As illustrated in Fig. 1, the process starts with entering a Malay input sentence. Our input data is not receiving a paragraph or single word. After the Malay input sentence is entered, the process of tokenising will execute. This process is to filter any unrecognised symbol or punctuation in the a text. All tokenised words will be assigned with Malay Part-of-Speech (POS) via executing the process Malay Part-of-speech(POS) Tagger module. After the words have been tagged, the next process called as a dependency based-words classifier module is executed. This module will do analysis by performing the process of recognising pattern matching structure with the tagged words from an input sentence. By using this module, the structure of a Malay input sentence, is segmented into four main phrases, such as *"Frasa Nama(FN)"* (Noun Phrase), *"Frasa Kerja(FK)"* (Verb Phrase), *"Frasa Adjektif(FA)"* (Adjective Phrase) and *"Frasa Sendi Nama(FS)"* (Prepositional Phrase). These group of phrases of a Malay sentence is determined by searching the rules in the rule-based phrase database. The accuracy of the pattern matching is determined by the number of collection rules extracted from a rule-based phrase database. After discovering the rules which match the phrases inside a Malay sentence, the process will continue to consummate a module for context-based similar word meaning detection. This module will decide the context of a word in a Malay text. The decision context of a word inside a sentence will use regulation *"inti"* (head) and *"penerang"* (modifier) as discussed in [4,5,9]. A Synonym contextual knowledge base will store the synonym word lists based on relevant context obtained from pertinent analysis sentences in Malay. The details of the process is shown in Fig.1.

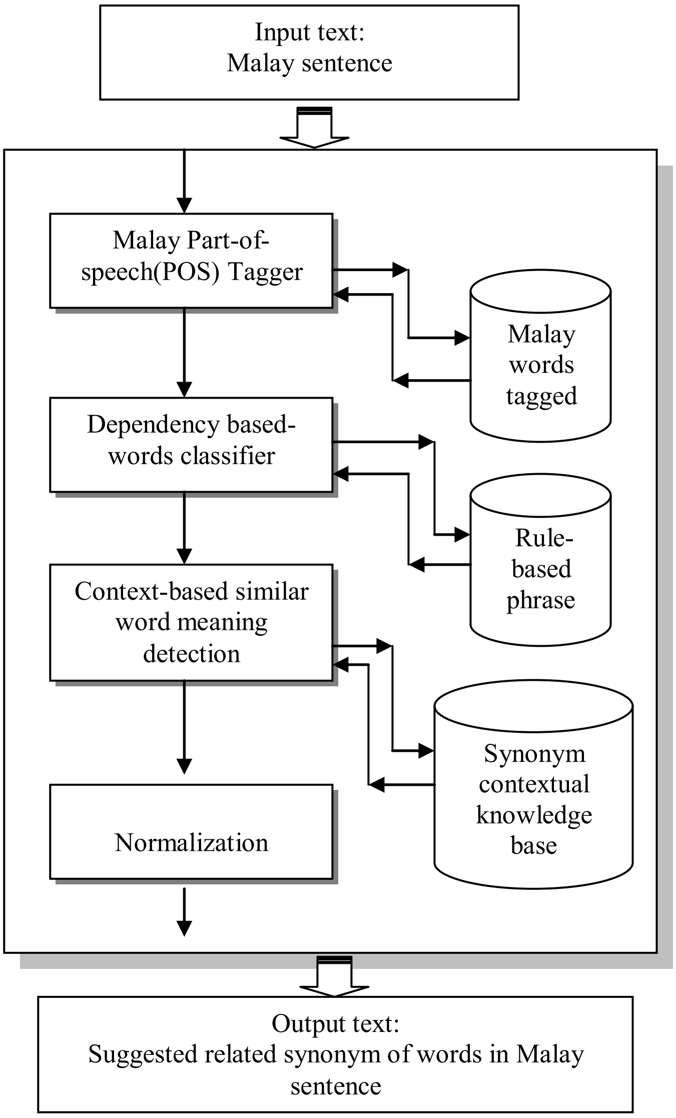

Input text:
Malay sentence
Malay Part-of-speech(POS) Tagger
Malay words tagged
Dependency based-words classifier
Rule-based phrase
Context-based similar word meaning detection
Synonym contextual knowledge base
Normalization
Output text:
Suggested related synonym of words in Malay sentence

Fig. 1 System flow of the process

A. *Malay Part-of-speech(POS) Tagger*

The Malay POS tagger was designed to allot POS for each word in Malay sentence. To construct Malay words with POS, we by hand select them from a Kamus Dewan Bahasa dan Pustaka (DBP) [3]. Currently, we have successfully compiled almost 18,135 words associated with their proper POS. Table 1 depicts the total number of words composed together with their own POS category.

TABLE 1
POS TYPES IN MALAY LANGUGE

Part Of-Speech (POS)	Total number of words
Kata Nama (KN)	6097
Kata Kerja (KK)	2536
Kata Adjektif/Sifat (KA)	1623
Kata Sendi (KS)	1407
others	6472

All the words tagged originates from a general domain data set. Below is a Table 2 depicts the list of some examples Malay POS tag that will use inside our data.

TABLE 2
POS TYPES IN MALAY

Part Of-Speech (POS)	Description	Example
KN	*Kata Nama*	*bangunan, kucing, sungai, etc.*
KK	*Kata Kerja*	*bermain, menulis, diambil, makan, etc.*
KA	*Kata Adjektif/Sifat*	*besar, cantik, rendah, tinggi, etc.*
KS	*Kata Sendi*	*di, ke, dari, dalam, etc.*
KBIL	*Kata Bilangan*	*satu, dua, empat, etc.*
KH	*Kata Hubung*	*yang, dan, atau, etc.*
KB	*Kata Bantu*	*masih, sedang, pernah, telah, sudah, akan, belum, harus, mesti, boleh, etc.*
KP	*Kata Penunjuk*	*itu, ini*
KAR	*Kata Arah*	*atas, tengah, bawah, antara, sisi, penjuru, segi, samping, utara, luar, etc.*

Based on the Malay POS tag stored in our Malay words tagged database, we can produce output for Malay sentence having POS alongside with their right word grammatical structure order. Fig.2 is an example of the Malay sentence before and after performing a tagging process.

Before: *"Pelajar itu sedang mengulangkaji pelajaran."*

↓ Tagging process

After: *"Pelajar[KN] itu[KN] sedang[KB] mengulangkaji[KK] pelajaran[KN]."*

Fig. 2 Example of Malay POS tagged inside a Malay sentence

The output from this module acts as a new input for the process called Dependency based-words classifier.

B. Dependency based-word classifier

The initial idea started with observing and understanding the pattern grammar structure of words in Malay sentence. As discussed in [4,5,9], all the sentences in Malay language were built based on the accumulation of four word categories, for instance "*Kata Nama*" (Noun), "*Kata Kerja*" (Verb), "*Kata Sifat*" (Adjective) and "*Kata Tugas*" (Function word). With the various combination of either using one or more of the above word categories, it can assist us to model a complete sentence structure in Malay language. Through thorough study and research on dependency structure in Malay sentences [4,5,9], we can clearly formulate series of standard rule-based structures. This is beneficial for handling and managing the possibility of different kinds of input sentences or phrases in Malay. Fig. 3 illustrates an example on the above discussion.

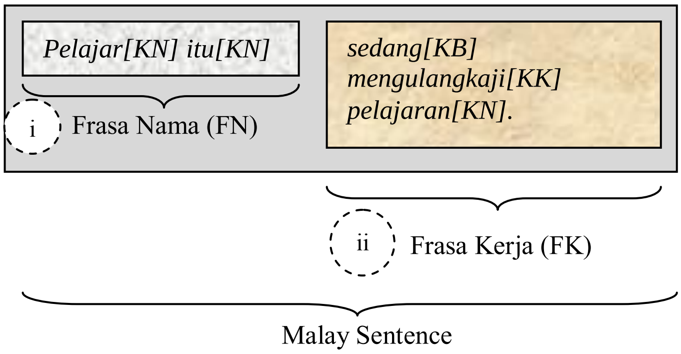


Fig. 3 Example of two phrases in a Malay sentence *"Pelajar itu sedang mengulangkaji pelajaran"*

Based on the findings, we can generate more related rules by taking examples of Malay sentences extracted from Kamus DBP [3]. Currently, we have successfully collected and analysed 50 short sentences in Malay to obtain the standard pattern rule-based structures to solve the problems defined. The more examples in Malay sentences study, the more standard rule-based structure can be produced. This is very important for us in determining the accuracy and coverage in producing the synonym of word lists based on context of a word in a Malay sentence. However, additional involvement from Malay language expert/linguist is very important in order to clarify the list of standard rules that should be in our approach.

Table 3 shows a few examples of the rule-based formulated. All the Malay phrases collected in these examples is from [9].

TABLE 3
RULE-BASED STRUCTURE

Phrases (rule-based)	Example
FN → KN	*FN → Ahmad, etc.*
FN → KN + PN	*FN → Kawasan + itu, etc.*
FN → KN + KN	*FN → gunting kain* *FN → alat pemadam* *FN → tudung kepala, etc.*
FN → KBIL + KN + (KN) + (KA) + (PN)	*FN → seorang pengakap* *FN →sebuah panggung wayang* *FKN →dua orang murid pintar itu, etc.*
FN → KN + KS + KN	*FN → rumah di bandar* *FN → surat dari kampung, etc.*
FN → KN + KK	*FN → wad bersalin* *FN → bilik mandi, etc*
FN → KN + KA	*FN → kotak hitam* *FN → kemalangan ngeri, etc.*
FK → KB + KK + KN	*FK → sedang menyanyi lagu* *FK → sudah makan nasi, etc.*
FK → KK + KN	*FK →memancing ikan* *FK → menendang bola, etc.*
FA → KN + KS	*FA → Dia gemuk*
FA → KN + KP +	*FA→ Buku itu terlalu tebal*

(penguat) + KS	

Regarding the above examples, we can perhaps detect the context of a word inside the phrase related with a word input. As noted in [4,9], every phrase in Malay sentence have *"inti"*(head). The word appearing as *"inti"*(head) is most significant compared with other words that signify as *"penerang"*(modifier), specifically in perceiving the word in Malay sentence, whether is in a context of a word or as others. Fig. 4 is an example of how the concept *"inti"*(head) and *"penerang"*(modifier) can be positioned into a sentence *"Pelajar itu sedang mengulangkaji pelajaran."*.

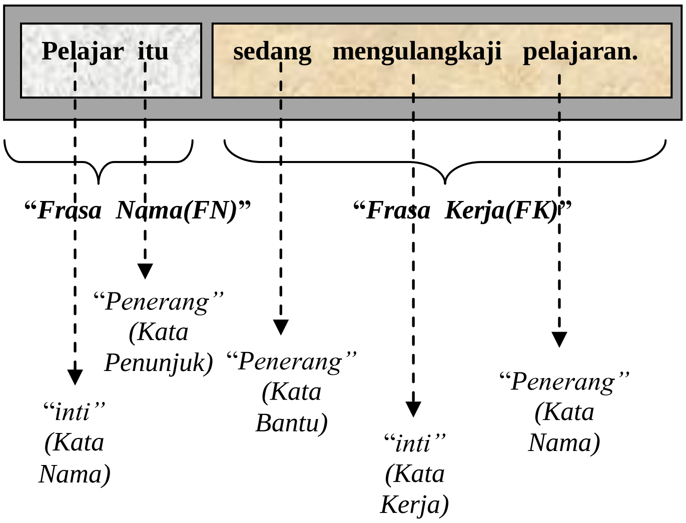


Fig. 4 Example of *"inti"* and *"penerang"* position in a Malay sentence

As referred from Fig. 4, the phrase of "*Pelajar[KN] itu[PN]*" is identified as "*Frasa Nama(FN)*" which follows the rule FN → KN + PN. Regarding this rule, the *"inti"* should be in word *"Pelajar[KN]"* while *"penerang"* is in word *"itu[PN]"*. The same concept apply for the rest of the words. For instance, the phrase of "*sedang[KB] mengulangkaji[KK] pelajaran[KN]*" is known as "*Frasa Kerja (FK)*" which matches the rule "*FK → KB + KK + KN*". Inside this "*Frasa Kerja[FK]*" there is also "*inti*" , such as in word *"mengulangkaji[KK]",* while the rests is *"penerang"* such as in word *"sedang[KB]"* and *"pelajaran[KN]"*. According to [4], the "*inti*" in each Malay sentence cannot be eliminated as compared with *"penerang"*.

Based on the above study, we can discover the context of word input inside each phrase with using the regulation *"inti"* and *"penerang"* seated into the sentences. However, more extensive research in this area is expected in finding more relevant approach or technique, which can assist to contribute into our further research work.

C. Context-based similar word meaning detection

The purpose of this module is to detect the phrases which may be present in a Malay sentence. The detection of phrases is depending on the list of rule-based phrase structure available from the rule-based phrase database. Fig. 6 shows the detection of the context in relation with the input word that appears in the phrase.

In reference to Fig. 6, the following analysis can be made:

Word input/search entered : *"mengulangkaji"*

Possibility of the context word detected:

a) First phrase detect : *"Pelajar itu"*
Context word detect : *"Pelajar"*
b) Second phrase detect: *"sedang mengulangkaji pelajaran"*
Context word detect: *"pelajaran"*

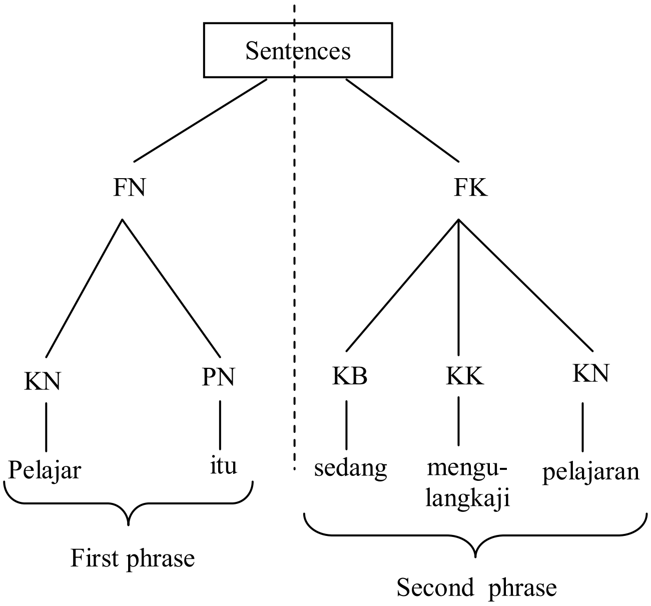


Fig. 6 Example of a tree structure for detecting contextual synonym in a Malay sentence *"Pelajar itu sedang mengulangkaji pelajaran"*

From the above analysis, if word *"mengulangkaji"* is selected as a word input, the process will be a self regulating checking mechanism and picks up the second phrase beside the first phrase to produce a result. This is because the location of word *"mengulangkaji"* was placed within *"Frasa Kerja (FK)"* rather than *"Frasa Nama(FN)"*. In addition, there are other adjacent to produce complete meaning for the phrase. In this case, we can determine the context of the word *"mengulangkaji"* by seeking them from a synonym contextual knowledge base database.

IV. CONCLUSION AND FURTHER RESEARCH

The process flow described is very important for us in obtaining an acceptable solution for the problems addressed. In earlier stage of comprehension and preparing data for our requirements, we have been pointed with a few worthwhile books such as [4,5,9] to get primary understanding on the essential requirements, specifically in studying syntax and grammar structure in Malay language. Aggressively sharing and discussing research work from other researchers also is very gainful to look with the current issues on topic research related.

In our further research, we have to address on the following main subject matters:

- To construct more words in Malay alongside with their POS tag.
- To identify and formulate a list of standard rules from selected Malay sentences, which can react with a variable to hold the words.
- To form and strengthen the entire process involved in our research work, especially focus more on the algorithms, formulas and data representations.
- To construct list of synonym groups based on context of a word into our database.
- To examine the effectiveness of the concept and approach applied as a problem solving in our study.

REFERENCES

[1] A. Budanitsky and G. Hirst, *Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures*, Proc. Workshop WordNet and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for Computational Linguistics, June 2001.

[2] A. Budanitsky, *Lexical Semantic Relatedness and Its Application in Natural Language Processing*, Technical Report CSRG-390, Dept. of Computer Science, Univ. of Toronto, Aug. 1999.

[3] Arbak Othman, Nik Safiah Karim, *Kamus komprehensif Bahasa Melayu*, Cetakan kedua, Penerbit Fajar Bakti Sdn.Bhd, 2006.

[4] Addullah Hassan, *Tatabahasa Bahasa Melayu*, 4th edition, PTS Publications & Distributors Sdn. Bhd. 2004.

[5] Abdullah Hassan, *Linguistik Am*, 10 edition., PTS Profesional Publishing Sdn.Bhd. 1992.

[6] D. Lin. *Automatic retrieval and clustering of similar words.* In Proceedings of COLINGACL,1998, pp. 768-774.

[7] Dekang Lin. 1994. *Principar—an efficient, broad coverage, principle-based parser*. In *Proceedings of COLING–94*, pages 482–488. Kyoto, Japan.

[8] Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiro Toyama. 2006. *Selection of Effective Contextual Information for Automatic Synonym Acquisition. Proc. of the 21st* International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 353-360.

[9] Nik Sapiah Karim,Farid M.Onn,Hashim Haji Musa,Abdul Hamid Mahmood, *Tatabahasa Dewan* , 3rd edition (cetakan kelima), Dewan Bahasa dan Pustaka (DBP), 2010.

[10] Pado, S. and Lapata, M. (2007). *Dependency-based construction of semantic space models. Computational Linguistics*, 33(2):161–199.

[11] Pereira, F., Tishby, N., and Lee, L. (1993). *Distributional clustering of english word*. In *Proc. of ACL 93*, pages 183–190.

[12] Yuhua Li, Zuhair A. Bandar, and David McLean, *An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources*, IEEE transactions on knowledge and data engineering, Vol.15, No.4, July/August 2003 871.

[13] Zeng Xian-mo. (2007). *Semantic relationships between contextual synonyms,* US-China education review, ISSN1548-6613, USA, Vol.4, No.9 (serial No.34).