Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Special Issue

Available online at http://www.jfas.info

EVALUATION ON KNOWLEDGE EXTRACTION AND MACHINE LEARNING IN RESOLVING MALAY WORD AMBIGUITY

M. F. Yahaya[1,*], N. A. Rahman[1], Z. A. Bakar[2] and H. Hasmy[3]

[1]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, Shah Alam, Selangor, Malaysia

[2]Faculty of Computer and Information Technology, Al-Madinah International University, Shah Alam, Selangor, Malaysia

[3]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA, PasirGudang, Johor, Malaysia

Published online: 17 October 2017

ABSTRACT

The involvement of linguistic professionals in resolving the ambiguity of a word within a particular context will produce a concise meaning of the words that are found in the lexical knowledge based collection. Motivated from that issue, we employed lexical knowledge and machine learning approach which includes the integration of data or/and information from the lexical knowledge based, that is Malay collections which linked to the ambiguous words. We used the most open class word and removed the stop words from the targeted sentences. Experiments have been conducted with and without lexical knowledge on 50 ambiguous words. The Word Sense Disambiguation (WSD) method is determined by machine learning, corpus based approaches namely Malay-Malay corpus and English-Malay corpus. The results show that the proposed method has improved the precision in resolving ambiguity.

Keywords: ambiguity; lexical knowledge; machine learning; Malay word.

Author Correspondence, e-mail: bustaguys@yahoo.com

doi: http://dx.doi.org/10.4314/jfas.v9i5s.10

1. INTRODUCTION

Ambiguous words can have multiple meanings [1]. For example the Malay word "*semak*" has several meanings such as to check, confuse, or having the same mother. Getting the right meaning of any ambiguous word is easy for human, but developing Natural Language Processing (NLP) system for machine is complicated [2]. However, this can be overcome by incorporating knowledge that identifies the true word of the uncertain word or called Word Sense Disambiguation (WSD) [3]. For WSD, a given a collection of words, a classifier is applied to produce two types of knowledge sources that distinguish senses of the words. The first is the corpus that is either not labeled or annotated with word senses. The second type is dictionary that can be machine readable dictionary, and thesaurus [2]. Without knowledge sources, it is hard either people or machines to recognize the correct sense. There are several WSD techniques ranging from knowledge or information based, either supervised or unsupervised techniques. Supervised or unsupervised techniques are depend in corpora prove [3]. Supervised technique has few annotation and has large annotated corpus [4].

2. RELATED WORK IN KNOWLEDGE EXTRACTION AND WORD SENSE

Knowledge based approach uses knowledge resources such as Wordnet and it is also referred as dictionary based approach [5]. To get the right sense, information construct depends in light of the word references [6] and proposed WSD with conceptual density method. This method selects words depend round the reasonable separation of the vague word and setting words that are connected [7] employ a selection inclination strategy. This strategy is used to find the likely relationship between word classes; easiest compute on words to the word's connection are recurrence check. Cover solely methodologies such Lesk. Lesk Extended is absolutely abased across the coordinating of word and settings words [8]. However, this approach is determined by the dictionaries and restricted in receiving the common sense knowledge.

3. METHODOLOGY

To assess the adequacy of question interpretation approaches in idea based machine learning system for Malay-English language pair, we conducted a progression of analysis utilizing the Malay-English arrangement corpora. The Malay-Malay corpus and English-Malay corpus as

in Fig. 1.

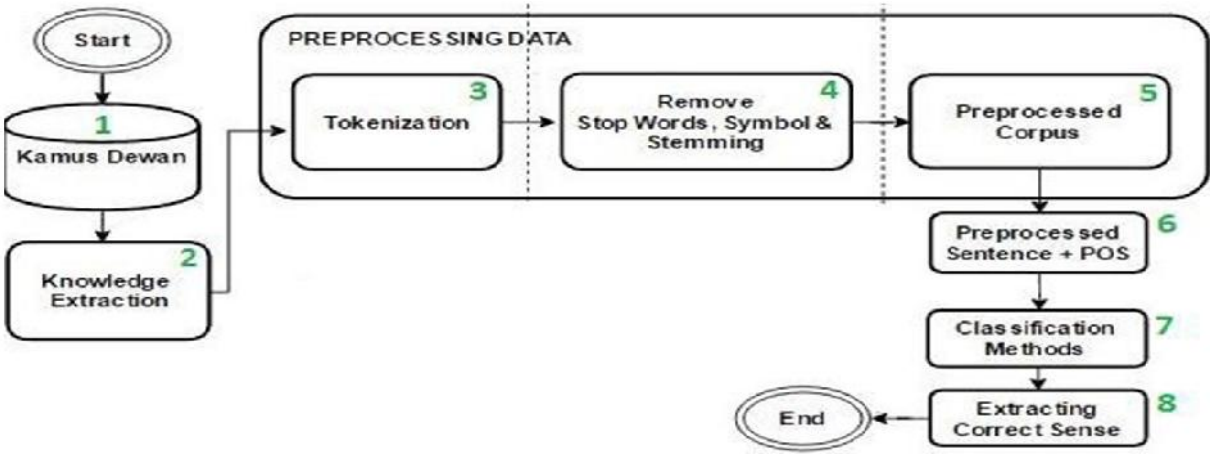

Fig.1. Overview in overall development conducting an experiment

3.1. Kamus Dewan (Text Corpora-Based)

For investigation reason, rundown of ambiguous words were constructed physically, comprises of legitimate names, for example, people name, verb, noun and events. There were 250 sections in Malay and English ideas word in view of 50 inquiries made for this trial. These arrangements of ambiguous words were utilized as a document and query interpretation handle.

3.2. Preprocessing

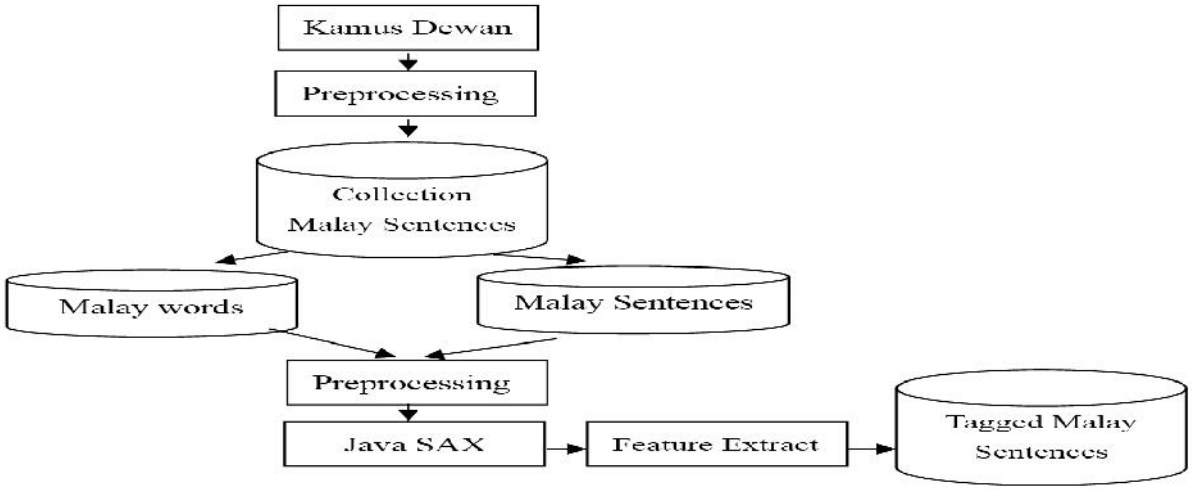

Fig.2. Construction of Malay words and sentences

Firstly, all words that are pertinent to ambiguous words were retrieved from [9], the Dewan Bahasa Pustaka(DBP) online dictionary. In this phase, words are used from different collections to seek the potential of ambiguous words. Fig. 3 shows the word *'semak'* submitted to the DBP online dictionary. In order to identify the ambiguous words, Fig. 4

shows that it has retrieved the answer from, KamusDewan offline dictionary. KamusDewan corpus is a comprehensive system that is manually constructed by the editor. KamusDewanBahasaPustaka uses JavaScript as a comprehensive hierarchical index of topics with different editions of the dictionary.

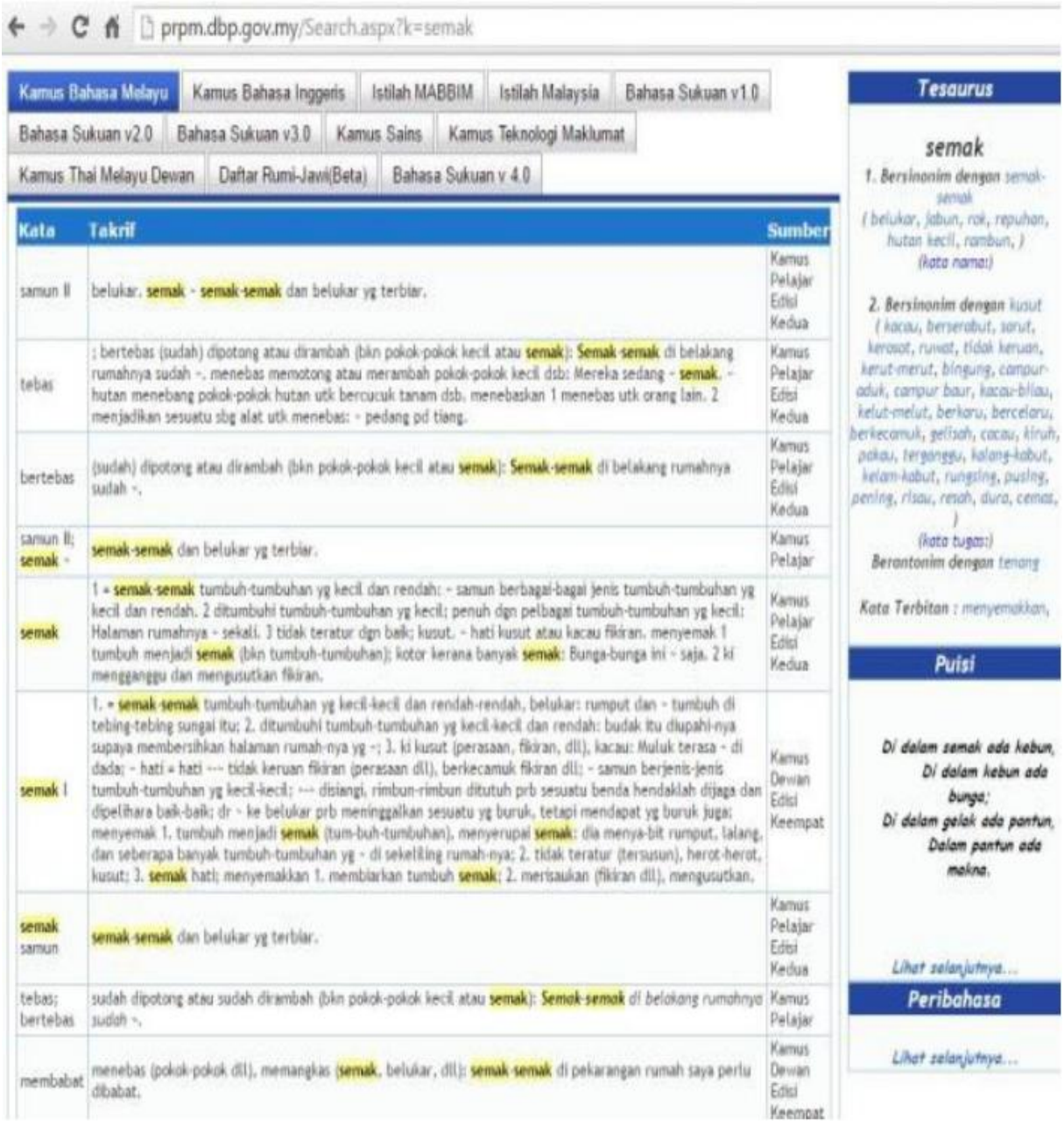

Fig.3. Output from DBP online dictionary for the word '*semak*'

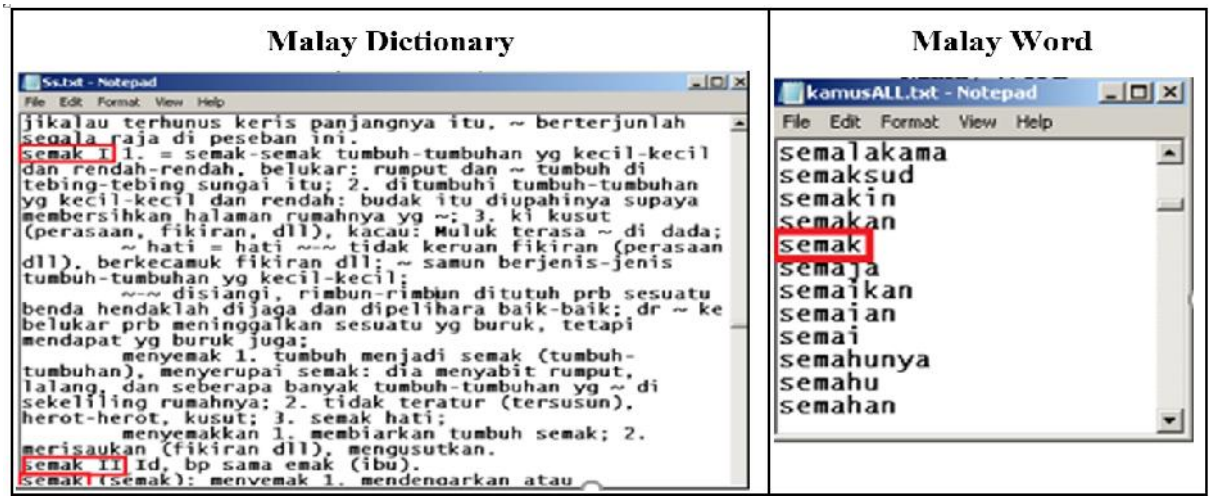

Fig.4. The word‘*semak*’ and sentences retrieved from the KamusDewan corpora

3.3. Malay Ambiguous Words: *“Semak” , “Sepak” , “Rendang” , “Perang”*

Concept-based lexical: In concept-based lexical knowledge, verb, adjectives, thing expressions are mapped into the ideas they speak to. In these methodologies, an archive is spoken to as an arrangement of ideas. Fig. 2 shows document and words involved eleven main tasks: KamusDewan, knowledge extraction, preprocessing data includes tokenization, remove stop words, symbol, preprocess corpus and features extraction includes target sentence, remove stop words, symbol and stemming, wordnet, classification methods and extracting correct sense.

3.4. Malay Words/Malay Sentences

Currently, most of Malay bilingual dictionaries did not provide detailed and clear information to help user in translating words. Therefore, this paper provides a bilingual dictionary with an explanation as follows:

Example 1: A word *‘semak’*, there are two homonyms, a noun *‘semak’* and a verb *‘semak’*. *‘semak’* is used sometimes as a noun and sometimes as a verb. The meaning of the word *“semak”* is used for checking something. The sentences are used to test the word in training set. The example sentences are as below:

1. *“Semakayat yang dibina, pastikanstrukturnyabetul”*. →Checking behaves more like a “verb”.
2. *“Tebaskanlahsegalasemakini”*. →Bushy behaves more like a “noun”.

Therefore, the meaning of checking it can be interpreted as bushy too. It has more than one tag.

Example 2: The word *"sepak"* can either interpret as slapping or kicking. The example sentences are as below:-

1. *"Nina sepakSitikeranamencuri".* → Slapping
2. *"Ali sepakabudengankuatsehinggakakinyabengkakteruk"* . → Kicking

Therefore, the meaning of slapping it can be interpreted as kicking too. It has more than one tag.

Problem found: Translation; how to translate in details: there are three methods that has been proposed at this phase: (1) We adopted query interpretation stop words which naturally expelled from the English using all words from KamusDewan online or offline database (2) Translate only the query word that is ambiguous with stop word using translation lexicons (3) Translate using WordNet bilingual dictionary.

The preliminary analysis is considered based on the motivation by the fact [10]. The result given in most of the translators is not accurate. For that reason, this research will help the user to reduce misinterpretation for the words found in the corpus. From KamusDewan corpus, an ambiguous word found such as word in Example 1 the word *"semak"*, translated from Malay-Malay KamusDewan dictionary or Malay-English from Wordnet, the word *"semak"* has identified as ambiguous with interpretation as "bushy" and *"sepak"* as its second interpretation competitor, despite the fact that *"semak"* is most habitually utilized interpretation for "checking".

Translate only query word that found in kamusdewan online dictionary. Then again, the second method is spurred by the reality method was motivated by the fact [11] translate only the query words that is ambiguous from Malay-Malay KamusDewan dictionary or Malay-English from Wordnet, stop words were automatically removed from the English translation, one can include all the possible ambiguous word in the target sentences. There were two results can be obtained from this methods: (1) improving the possibility of understanding for each sentences from KamusDewan dictionary searching output; if all interpretation incorporated into the queries interpretation having the same meaning or (2) decreasing the performance in KamusDewan searching output; if the results given as high as accurate assumption translations [12].

There were two kinds of word references being utilized as a part of this trial, WordNet and

KamusDewan online/offline dictionary. Both reports and query are critical viewpoint in machine learning. In machine learning, if a word retains many translations, it is hard to figure out which target expression is the best contender for a source expression can have numerous interpretations and diverse settings prompt to different translations [24]. A basic answer for this issue is to isolate the word arrangement, interpretation rules choice, reordering and structure forecast, dialect model or joint interpretation expectation, language model or joint translation prediction. Example 1 indicates English to Malay dialect interpretation of Query 1(semak) and Example 2 demonstrates English to Malay dialect interpretation of Query 2(sepak).

3.5. Knowledge Extraction

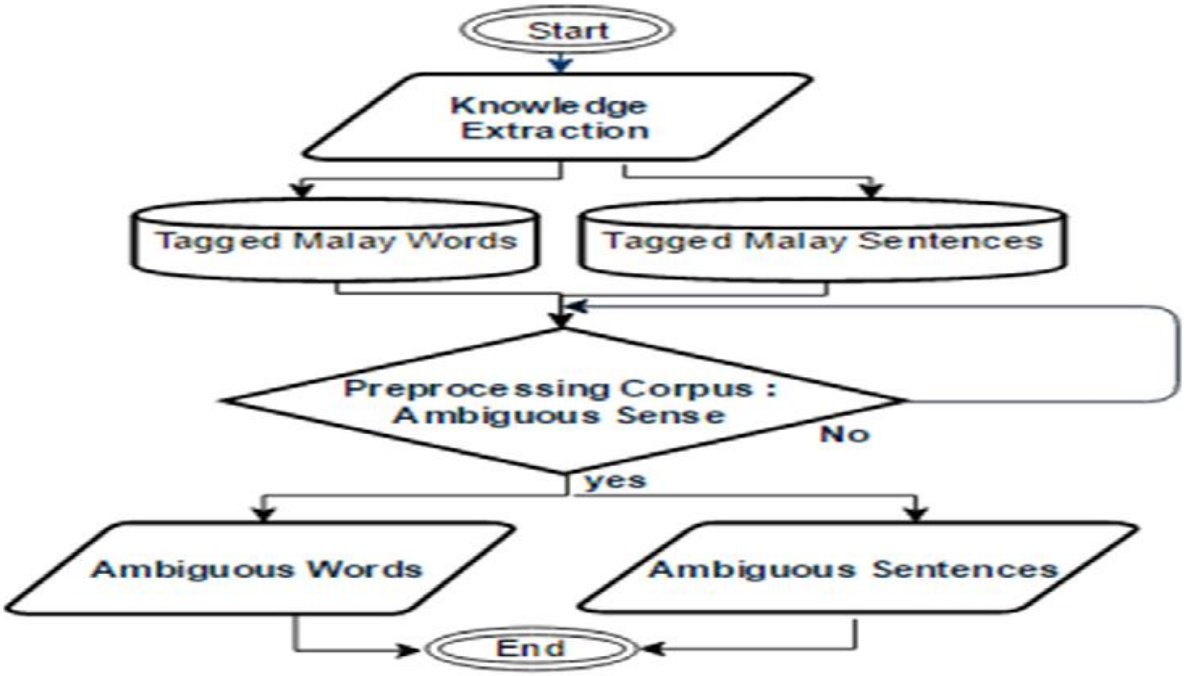

Fig.5. Input process output (ambiguous sense)

Fig. 5 shows the input and output for ambiguous sense. In order to test the words ambiguity, firstly, extract the words from the KamusDewan corpus. Then, once the ambiguity of the words were identified, tag the potential ambiguous Malay words to assign the correct tagged for each word and sentences. Afterwards, the correct tagged that assigned to the words and the sentences will identify the correct sense. As a result, it will bring the group of corrected ambiguous words and sentences [13]. For an idea Ci made out from words from ‘n’ that is recurrence on an inquiry equivalents for quantity events from an idea themselves, also from

whole sub ideas itself. Calculation format as below.

$$cf(c_i) = count(c_i) + \sum_{sc \in sub(c_i)} \frac{length(sc)}{length(c_i)} count(sc) \quad (1)$$

where length ci speaks to the quantity of words that shape ci and sub ci as an arrangement from whole conceivable sub ideas where is gotten from ci: ideas of n-1 words from ci, ideas of n-2, and every single expression of (ci).

3.5. Tokenization

Removing all the stop words, symbols and stemming to the user queries before tokenization process take in place [14]. Tokenization will produces queries catchphrases in view of significance idea words. All pertinent idea words existed in the question will be dealt with as one term. This rundown incorporates all relational words, pronouns and articles (which were at that point evacuated utilizing the morphological analyzer); normal stop words, for example, *"adapun"*, *"agar"* and *"betul"*.

For instance of our preparing consider the Malay queries on Example 1*: "Semakayat yang dibina, pastikanstrukturnyabetul"*. After morphological examination and stop word evacuation, this question gets to be "*semak*", *"bina"*, *"pasti"*, *"ayat"*. Each of these words is then turned upward in a Malay-English word-to-word lexicon, which contains the accompanying interpretations:

Table 1. Queries in Malay-English (removing stop words)

Semak	Bina	Pasti	Struktur	Ayat
check revise runthrough	build construct erect	surecertain definite positive reliable	structuresystem	sentence

In the Malay dictionary, some of the query word was lemmatized (e.g., an original word *"strukturnya"* was decreased to *"sktruktur"* in our test adaptation). In the event that a source word was not found in the lexicon, then the first source word was held in the interpretation. Document writings were tokenized into single words in tokenization prepare [15]. At that point, we consolidate a solitary word to make numerous word phrases. In idea recognizable

proof, the expressions made in past stride were distinguished as ideas based. The dictionary gaze upward was happen after tokenization assignment to interpret root words and compound words. Tokenization makes compound words interpretation conceivable.

Example 3: The word *"rendang"* is used to describe about a dry fried type of traditional Malay dishes or to describe about the leafy and small tree that gives shade. The example sentences are as below:

1. *"MakMinahmemasakrendangayamuntukdihidangkankepadatetamukendurikesyukuran."* →Dry fring
2. *"Apabila kami berhentikeranapenatberlumbalari, kawan-kawanmengajakuntukberehat di ataskerusipanjangterletak di bawahsebatangpohon yang rendang."* → Leafy

Example 4: The word *"perang"* can be used for brown hair and going for war as below:

1. *"Kebanyakan orang Amerika Latin mempunyairambutberwarnaperang."* →Brown
2. *"Ramai orang Islam yang matiberjihadakibatperang."* →War

Frequency Counts of Ambiguous Malay Words

Table 2 shows the frequency counts retrieved from over 50000 words corpus taken from the KamusDewan online dictionary. Only 50 ambiguous words were extracted from the KamusDewan. The highest number of occurrences would be the most ambiguous words used for each sentence. The ability to reduce the ambiguity is the main focus in this paper.

Table 2. The highest numbers of occurrences in KamusDewan corpus

Word	Frequent
semak	321
sepak	277
redah	276
redak	273
sepai	266
bela	265
redang	255
cetek	238
sela	224

sampuk	219
keras	219
jalan	211
selak	210
masak	209
gantung	207
buah	206
bekas	206
salak	205
tampar	204

3.6. Removing Stop Words, Symbol and Stemming

This procedure additionally lessens the content information and enhances the framework execution. Each word reference of KamusDewan deals with these words, which are a bit much for word sense. The stop words are not useful for word sense.

3.7. Preprocessed Corpus and Part of Speech (PoS)

Preprocessing is an imperative undertaking and basic stride in word sense [16] in Information Retrieval (IR) and NLP. In our research, the word sense-data preprocessing utilized for removing fascinating and superfluous words and learning from KamusDewan contents. IR is basically a matter of choosing which records in a gathering DBP ought to be recovered to fulfill a client's requirement for data entered [17]. The DBP provide to the client's requirement for data is spoken to by a question or profile and contains at least one inquiry terms in addition to some extra data, for example weight of the words. Henceforth, the recovery choice is made by contrasting the terms of the inquiry and the list terms (critical words or expressions) showing up in the archive itself.

Target sentences were recognized at this stage, the principal units go to all further preparing stages, from examination and labeling segments, for example morphological analyzers and grammatical form taggers; through applications, for example data recovery and machine interpretation frameworks. It is a collection of exercises in which in KamusDewan are pre-handled. Since the content information regularly contains some unique configurations like number organizations, date groups and the most widely recognized words that far-fetched to

help text mining, for example relational words, articles and professional things can be dispensed with [18].

KamusDewan[9] is in an unstructured format. In order to extract information from the kamus, the sentences are transformed into a structured format. These sentences are further process to remove the stop words and are stemmed into their root words. After the preprocessing step, features are extracted to train a classifier and to classify the correct sense without using lexical knowledge and using lexical knowledge as shown in Table 3 [19].

3.8. Wordnet

WordNet is available for the public and it is a lexical database produced by Princeton University [20]. In WordNet 3.0, about 200000 word senses into 117777 SynSets which are gatherings of equivalent words. This implies there are 117777 unique definitions that are available to get the correct specification of sense and semantic relations for each word [21]. WordNet is indeed rich in information and it is a standout amongst the most commonplace apparatuses for word sense disambiguation [22].

3.9 Classification Task (Method)

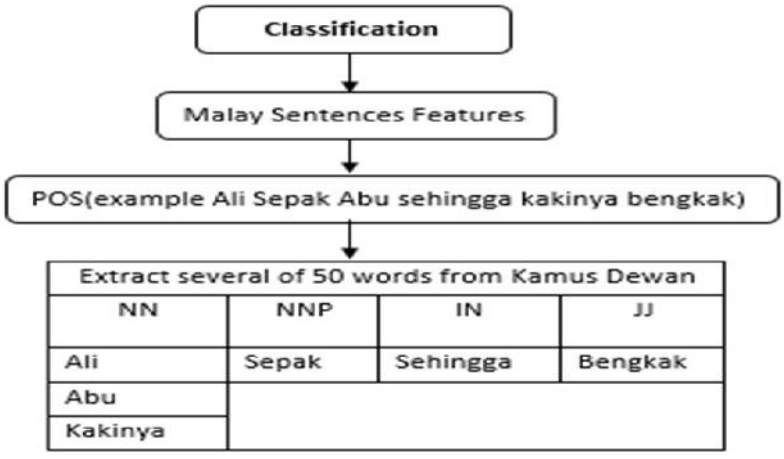

Fig.6. Classification of the words from target sentences

In arrangement methodology, subsequent to evacuating capacity words, just to justify the words content, starting at target locating dubious word at sentence. A shot in higher than 1 word uncertainty from the goal sentence. It needs for searching questionable sense word assistance by Wordnet. Furthermore, stride to play out the procedure in pre-handling of target sentence steps. Next, apply JAWS library methods.Therefore, the component is contrast with

the elements in preparing information. To check increment in every component information are discovered as well as information in target highlight, also a preparation includes [23].

3.10. Extracting Correct Word Sense

As approach is half and half one and we are utilizing probabilistic measure. Parameters in the probabilistic WSD are: Pr(s) i.e likelihood of sense and Pr(Vwi |s) i.e. likelihood of highlight w.r.t. specific sense Pr(s)= count(s,w)/count(w) and Pr(Vwi |s)= count(Vwi ,s,w)/count(s,w) The sense with the most noteworthy likelihood is return and alongside its sense id. This sense id is utilized to outline definition connected with that sense for the objective word in the sentence.

4. RESULTS AND DISCUSSION

As the results display in Table 4, it shows that the translation can be extremely setting delicate where it comes with a few outcomes. To start with, which means explanation is extremely delicate to the sorts of ideas/relations/learning present in the vocabulary. For instance, in the last case in Table 4, *"Perang"* in the setting was deciphered as "where individuals who have a chestnut hair" or "go fight". This elucidation happened, despite the fact that individuals are going war. This needs to do with the significance of relations in the trial of Malay Lexicon. In spite of the fact that those other individuals who have 'brown hair' were present, they were not appropriately associated with "going war". This proposes that importance is not just affected by what exist in the context sentence, it is vigorously impacted by what is truant from the specific situation, for example the non-appearance of a connection that ought to exist.

Table 4. Results from experiment conducted to determine the effects of active context of Malay words *"semak", "sepak", "rendang"* and *"perang"*

Context(AmbiguousWord)	**Top Interpretation(Scorein%)**
Semak (check)*periksa*	*Semakayatyangdibina,pastikanstrukturnyabetul*(25%)
Semak (bushy)*belukar*	*Tebaskanlahsegalasemakini*(35%)
Sepak (slap)*lempang,tampar*	*NinasepakSitikeranamencuri*(55%)
Sepak (kick) *tendang*	*Ali sepakabudengankuatsehinggakakinyabengkakteruk (60%)*
Rendang (dryfrying)*masakan*	*MakMinahmemasakrendangayamuntukdihidangkankepadatetamukendurikesyukuran*(30%)
Rendang (shady)*daunlebat*	*Apabilakamiberhentibermainbolasepak,kamiberehatdibawahsebatangpohonyangrendanguntukseketika*(40%)
Perang (brown)*warna*	*KebanyakanorangAmerikaLatinmempunyairambutberwarnaperang* (35%)
Perang (war)*bertempur*	*Ramaiorangislamyangmatiberjihadakibatperang*(45%)

We first apply Malay-Malay, Malay-English test corpus which is composed of 50 ambiguous words extracted from KamusDewan, and it is one of the most authoritative test corpus currently used to evaluate Malay WSD method. KamusDewan provides frequency counts from over 250 words of parallel sentences and its definition on a web http://prpm.dbp.gov.my/Search.aspx?k=. However, we found that the top 4 highest frequency obtained are *"semak","sepak", "rendang"* and *"perang"* from Malay-Malay corpus. We then supplemented the four ambiguous words (to classify the words on the right tags). We collected a total of 4 ambiguous words to test the ambiguous Malay-English Parallel Corpus

in WordNet.

5. CONCLUSION

We used KamusDewan and WordNet to evaluate our Word Sense result. The three values reflect the completeness, the accuracy and the effectiveness of the approach respectively. In this paper, we prepared a larger text set of full texts to evaluate our approach.

6. ACKNOWLEDGEMENTS

The research was funded by the Malaysian Government under Fundamental Research Grant Scheme (FRGS) (FRGS/1/2015/ICT01/UITM/03/1) in UniversitiTeknologi MARA, Shah Alam.

8. REFERENCES

[1] Eddington C M, Tokowicz N. How meaning similarity influences ambiguous word processing: The current state of the literature.Psychonomic Bulletin and Review. 2015, 22(1):13-37

[2] Higginbotham DJ,Lesher GW, Moulton BJ, Roark B. The application of natural language processing to augmentative and alternative communication. Assistive Technology, 2012, 24(1):14-24

[3] Zhao H, Kit C. Integrating unsupervised and supervised word segmentation: The role of goodness measures. Information Sciences, 2011, 181(1):163-183

[4] MontoyoA, Palomar M, Rigau G, Suárez A. Combining knowledge-and corpus-based word-sense-disambiguation methods.Journal of Artificial Intelligence Research. 2005, 23:299-330

[6] Zouaghi A, Merhbene L, Zrigui M. Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation. Artificial Intelligence Review, 2012, 38(4):257-69

[7] Yang CY, Wu S J. Semanticweb information retrieval based on the wordnet. International Journal of Digital Content Technology and its Applications, 2012, 6(6):294-302

[8] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using

WordNet. In International Conference on Intelligent Text Processing and Computational Linguistics, 2002, pp. 136-145

[9] DewanBahasadanPustaka. KamusDewan. 2010

[10] Ramteke S, Ramteke K, Dongare R. Lexicon Parser for syntactic and semantic analysis of Devanagari sentence using Hindi wordnet. International Journal of Advanced Research in Computer and Communication Engineering, 2014, 3(4):6345-6349

[11] Yahya Z, Abdullah M T, Azman A, Kadir R A. Query translation using concepts similarity based on Quran ontology for cross-language information retrieval. Journal of Computer Science, 2013, 9(7):889-897

[12] Sakurai N. The influence of translation on reading amount, proficiency, and speed in extensive reading. Reading in a Foreign Language, 2015, 27(1):96-122

[13] Zampieri M. A supervised machine learning method for word sense disambiguation of Portuguese nouns. Bulletin de LinguistiqueApliqueetGnrale-BULAG, 2010, 34:187-203

[14] Ayedh A, Tan G, Alwesabi K, Rajeh H. The effect of preprocessing on Arabic document categorization. Algorithms, 2016, 9(2):1-17

[15] Rehman Z, Anwar W, Bajwa U I, Xuan W, Chaoying Z. Morpheme matching based text tokenization for a scarce resourced language. PloS One, 2013, 8(8):1-8

[16] Zhong Z, Ng H T. Word sense disambiguation improves information retrieval. In 50th Annual Meeting of the Association for Computational Linguistics, 2012, pp. 273-282

[17] Zainudin I S, Jalaluddin N H, Bakar K T A. The use of corpus and frame semantics in a lexicography class: Evaluating dictionary entries. Procedia-Social and Behavioral Sciences, 2014, 116:2316-2320

[18] Fonseca E R, Rosa J L G, Aluísio S M. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. Journal of the Brazilian Computer Society, 2015, 21(1):1-14

[19] DeLong K A, Troyer M, Kutas M. Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. Language and Linguistics Compass, 2014, 8(12):631-645

[20] Miller G. WordNet-About us. WordNet, New Jersey: Princeton University, 2010

[21] Kilgarriff A, FellbaumC.WordNet: An electronic lexical database.Language,

2000,76(3):706-708

[22] Navigli R. A quick tour of word sense disambiguation, induction and related approaches. In International Conference on Current Trends in Theory and Practice of Computer Science, 2012, pp. 115-129

[23] Pal S, Pakray P, Naskar S K. Automatic building and using parallel resources for SMT from comparable corpora. In 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL,2014, pp. 48-57

[24] Bond F, Lim L T, Tang E K, Riza H. The combined wordnetbahasa. NUSA: Linguistic studies of languages in and around Indonesia, 57:83-100

How to cite this article:
Yahaya M F, Rahman N A, Bakar Z A, Hasmy H. Evaluation on knowledge extraction and machine learning in resolvingmalay word ambiguity. J. Fundam. Appl. Sci., 2017, *9(5S), 115-130.*

Table 3. Algorithm to classify correct sense

Steps	WithoutLexicalKnowledge	WithLexicalKnowledge
Training	• Extract features from corpus sentences • Classify the words with classifier extracted features. • For the motivations behind making and keeping up a computational vocabulary, it might be alluring to perform administered preparing on the dictionary to learn specific significance for Malay words	• Extract features from corpus sentences • Classify the words with classifier extracted features.
Disambiguation	• Select words across from the selected words • From the feature set will have to compare the ambiguous word which is integration of data source from their class • Compute a possibilities for every word sense	• Select words across the target word • Include world knowledge • Compare ambiguous word (also consider here the lexical expertise in ambiguous word) words using the word of feature set and that is integration of
Winner Sense	Just compare and assign the result directly of each sense to the highest sense	Need to compare and assign each sense to the maximum sense