Incorporating Knowledge Base in Unsupervised Approach of Word Sense Disambiguation of Malay Documents

Mohd Arizal Shamsil Mat Rifin and Mohd Pouzi Hamzah
School of Informatics and Applied Mathematics,
Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
arizalshamsil@gmail.com

***Abstract*—The problem of ambiguity in a text document or query is among the issues found in information retrieval. This problem occurs when a word has more than one meaning. The presence of ambiguity in a text or query will have a negative impact to the information retrieval process and the query expansion process. Addition of supplementary keywords in the query expansion process would be inaccurate without identifying the exact sense of the word. Ambiguous terms need to be disambiguated to avoid this problem. The process of identifying the proper sense is known as word sense disambiguation (WSD). The study of word sense disambiguation in text documents have been carried out by researchers worldwide. However, a study on this issue in the Malay language context is still insufficient. The proposed method is an adaptation of a famous unsupervised and knowledge-based method.**

***Index Terms*—Information Retrieval; Malay Text; Unsupervised; Word Sense Disambiguation.**

I. INTRODUCTION

Every word we speak and write is definitely a word that has different meanings [1]. Such words are called polysemy words. The presence of such words in a query will make the query or document ambiguous which means it cannot be interpreted accurately by computer leading to inaccurate results. Example of a word that has many meanings is the word "daki" in the Malay language. The first meaning of the word "daki" is climbing and the second meaning is dirt on the skin. Without identifying the right sense of this word, a query would be ambiguous and unclear. There are many benefits that could be gained if the problem of ambiguous terms can be resolved, particularly in systems such as the knowledge extraction, machine translation [2] and the information retrieval system [3]. Information retrieval systems such as Google or Yahoo search will become more powerful by removing the ambiguity from the queried term.

This paper is divided into five sections. Section I which is the introduction provides a general overview and the need for word sense disambiguation in a text document. Meanwhile, Section II describes previous research that has been done in the field of word sense disambiguation, categorized by type of word sense disambiguation technique. Section III presents previous researches that have been done on word sense disambiguation in Malay documents. This paper continues with the proposed method for Malay word sense disambiguation in Section IV and the conclusion in Section V.

II. RELATED WORKS

Word sense disambiguation (WSD) is a process to identify the exact sense for an ambiguous word. Based on [4], word sense disambiguation is the task to determine which sense of a word is correct in a particular context. Word sense disambiguation technique is grouped into three general groups which are supervised, unsupervised, and knowledge-based approaches [4]. Supervised approach is an approach that depends on large sense annotated data and machine learning algorithm to determine the sense of a word [1]. Unsupervised word sense disambiguation is an approach that is different or contradictory to the previous method. It is because this approach does not use a tagged corpus as a source of knowledge to do sense determination. This approach only needs to have raw annotated data to disambiguate the sense by using some kind of similarity measure [7]. The last approach of word sense disambiguation is the knowledge-based approach. These systems rely mainly on information drawn from lexical resources, such as dictionaries or thesauruses.

A. Supervised WSD

Based on [5], this approach has the highest performance and accuracy for word sense disambiguation. However, this approach is limited by the amount of sense annotated corpus for training models on all word types because the largest corpora contains only hundreds of thousands of annotated tokens [5]. It is also a big issue for supervised word sense disambiguation because the annotated corpus needs to be done by humans or experts in the linguistic field manually. Based on [1], researches have been done by several researchers around the world to explore the approach in the making of an automated sense tagged corpus for example by using a machine learning algorithm. Methods that use the supervised approach in determining the sense of a word are decision list, decision tree, naïve Bayes, Neural Networks, Instance Based Learning, Support Vector Machine, and Ensemble Methods which can be categorized into Majority Voting, Probability Mixture, Rank Based Combination, and AdaBoost [6].

B. Unsupervised WSD

Based on [1], these approaches can cluster word sense by not even referring to the sense inventory and tagged corpus which removes the limitation of the supervised method. However, this method is still second to the supervised method

as the best unsupervised word sense disambiguation can only manage to achieve around 70% precision and 50% recall [7]. In addition, the unsupervised word sense disambiguation can be categorized into several famous methods which are context clustering, word clustering, Co-occurrence Graph, and Spanning tree based approach [6].

Research has also been done by [8]. In this paper, the researcher proposed the word sense disambiguation method based on the unsupervised method by applying several unique concepts which include the one sense per collocation concept and the one sense per discourse concept. For the one sense per collocation concept, neighboring words in a query or document have strong and consistent inklings to the sense of an ambiguous word but affected by the order of the word from the ambiguous word [9], relative distance, and syntactic relation between terms. Besides that, the one sense per discourse is a concept which states that a word is extremely constant in one document, which means when one sense occurs in one document, it has a higher tendency that a similar sense will occur again in that document.

This algorithm works surprisingly well for an unsupervised approach, directly outperforming Schiitze's unsupervised algorithm by 96.7 % to 92.2 % when tested using a similar word. Furthermore, it is almost comparable to the result of the supervised algorithm for similar training contexts (95.5 % vs. 96.1%), and achieves higher performance in certain cases when using the one sense per discourse constraint (96.5 % vs. 96.1%). The finding of the study shows that the cost of a large sense tagged training can be left over to achieve accurate word sense disambiguation with low labor cost and reduced time consumption.

Research has also been done by Ivan Lopez-Arevalo in [4]. This study is about word sense disambiguation in a specific domain. This approach is done by identifying the major sense of ambiguous words from Wordnet. In addition, this method works by embedding two corpora which are domain-specific test corpus (contains target ambiguous words) and domain-specific auxiliary corpus (obtained by using relevant words from the domain-specific test corpus). This method consists of four key steps, which are (1) auxiliary corpus generation; (2) related features extraction (from the auxiliary corpus); (3) test features extraction (from the test corpus); and (4) features integration. This approach has been tested on domain-specific corpora (Sports and Finance) and on one balanced corpus, BNC. However, this approach showed some restrictions when dealing with the general-domain corpus but the obtained results for domain-specific corpora were better compared to previous works.

C. Knowledge-based WSD

In this approach, the disambiguation process is done by using similarity matching with definition of the word from a lexical resource such as a dictionary and thesaurus. The most famous method that uses the knowledge-based approach for word sense disambiguation is the Lesk algorithm [10]. In this method, the correct sense is decided by measuring the similarity of an ambiguous word with the definition provided by the dictionary. Since this study is a first attempt for knowledge-based word sense disambiguation, the accuracy of this system is only at the 50-70% range on some short samples.

Kanika Mittal and Amita Jain in [3] proposed a method for word sense disambiguation by comparing and finding the similarities between the ambiguous term with another term appearing in the query and by providing the weight to the calculated similarity. Value weighting for the similarity calculation of a particular word will be given in a descending order based on their distance from the ambiguous term. The value of the aggregate equation based on the weight given will be calculated using the operator Ordered Weighted Averaging (OWA) for each sense of the ambiguous term. The sense that has the highest value similarity will be considered the most suitable to sense a particular ambiguous term.

Referring to the past researches, the previous three approaches have their own advantages and disadvantages. Based on [6], the advantages and disadvantages of these approaches are shown in Table 1.

Table 1
Comparison of the established approaches

Approach	Advantage	Disadvantage
Supervised	This approach is said to be better than the unsupervised and knowledge based approach.	These algorithms do not offer a good result for resource limited languages.
Unsupervised	This method is not restricted by the size of the sense annotated corpora.	This method is more daunting to undertake and has low performance compared to the other two approaches.
Knowledge-Based	Greater Accuracy	This algorithm depends on the intersection with a dictionary so its performance is highly influenced by it.

Based on Table 1, the general supervised approach can provide the highest accuracy compared to the other two approaches which are the unsupervised and knowledge-based approaches. However, this method has a limitation because it is highly influenced by the size of the human sense tagged corpus, which consumes more time and a lot of human effort to be done. Moreover, the unsupervised approach is the most reliable approach which provides high potential for word sense disambiguation as the accuracy of this approach could defy the supervised approach without being limited by the size of the human tagged corpus.

III. MALAY WSD

Besides the studies that have been done by researchers on word sense disambiguation in English, there are also studies that have been done on word sense disambiguation for the Malay language. Among them, studies have been made in [11]. In this study, a word prediction algorithm, n-grams, was used to disambiguate the sentence. Prior to this, the word prediction algorithm was applied in helping the disabled to use technologies [12]. This study is an experiment to find out whether the word prediction algorithm is suitable to be implemented in resolving the ambiguity in Malay documents. Table 2 is the example of the result from the Maximum Likelihood Estimation (MLE) [13] of the bigram and trigram produced in this research.

The next study related to the Malay word sense disambiguation is [14]. This study used the unsupervised and conceptual clustering approach to investigate the existing Malay NLP tools to build a learning taxonomy from Malay texts for the proposed ontology learning approach. The tools

are a maximum-entropy parser based on an open NLP package, a word sense tagger, and a parser based on a polar grammar. A case study approach is adopted in this study and deemed suitable for an exploratory research. However, the result of this study shows a lower recall and precision for each NLP tool; nevertheless, this result does not prove that the unsupervised approach is unsuitable for Malay documents because the poor result may be due to several factors such as the texts being used in this experiment which are not original Malay texts but a translated text from the Hadith and Quran.

Table 2
Top 20 MLE for the Word "Madu"

Trigrams (madu, *, *)	Frequency	MLE
(madu, kepadanya, kemudian)	4	0.5000
(madu, beliau, bersabda)	3	1.0000
(madu, di, rumah)	3	0.7500
(madu, kemudian, orang)	3	1.0000
(madu, maka, aku)	3	0.7500
(madu, dan, aku)	2	0.1818
(madu, kepadanya, tapi)	2	0.2500
(madu, lalu, aku)	2	0.4000
(madu, adalah, al)	1	1.0000
(madu, atau, anggur)	1	1.0000
(madu, ayat, sebelumnya)	1	1.0000
(madu, bagaimana, itu)	1	1.0000
(madu, bernama, bit)	1	1.0000
(madu, bersama-sama, maka)	1	1.0000
(madu, biji, gandum)	1	1.0000
(madu, bila, beliau)	1	1.0000
(madu, dalam, bab)	1	1.0000
(madu, dan, al)	1	0.0909
(madu, dan, cangkirnya)	1	0.0909
(madu, dan, dibakar)	1	0.0909

The unsupervised method was once again the selected approach for the Malay WSD in [15]. In this paper, the researcher mentions that there are two traditional approaches for word sense disambiguation which are corpus-driven and learning-based and famously known as supervised and unsupervised methods. However, the unsupervised approach was chosen because it is not limited by a manually sense tagged corpora as in the supervised approach. Researchers have proposed a method using Cross-Language to reduce ambiguities in Malay-English Translation [9]. This method consists of four modules which are: word construction and extraction, word translation and computation, disambiguation, and evaluation. By translating an ambiguous word into an English word, the similarity of each sense of the ambiguous word with the English word is calculated by using several methods such as Path similarity [16], Lesk [10], Context Vector [17], and Vector pair [18]. This method can achieve quite a high accuracy at 78.79%. However, this technique is still not stable since the lowest accuracy is at 12.12%.

IV. Proposed Method

Based on previous research, the advantages and disadvantages for each approach are now clear. Supervised approach is seen as the most accurate WSD method but it is limited due to the bottleneck problem which is the size of the sense tagged corpus. Knowledge-based approach can also achieve a higher accuracy but since these algorithms are overlap-based, they suffer from overlap scarcity and its performance depends on dictionary definitions. Unsupervised approach has the right potential to be a good approach although its accuracy is just a little bit lower compared to the previous method. However, this approach has been the most selected approach by previous researchers as its potential is high and it does not have the previous two methods' weaknesses, which are the bottleneck and scarcity issues. Thus, the unsupervised approach is selected in this paper as the main part of the Malay word sense disambiguation method.

The proposed method goes through two main phases (as shown in Figure 1). The processes begin by identifying the number of ambiguous terms in a text or document by referring to the existing Malay Wordnet. Next, the step proceeds by identifying all the texts in the corpus that contains the selected ambiguous term. These texts are placed in the MySQL table containing the information as document id, term, and text that has the ambiguous term including its neighboring words. This step is important to identify the collocation term for the ambiguous term which will be done in the next process.

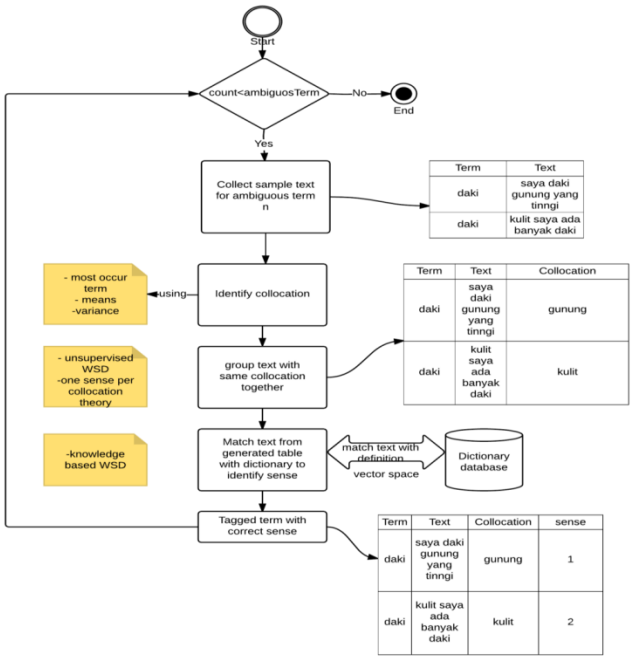

Figure 1: Proposed Malay WSD

In order to identify the collocation of a term, several techniques can be used which include calculating the frequency of the most simultaneously occurring terms with the ambiguous term and by using mean and variance [13]. By listing the frequency of the terms that occurs together with the selected ambiguous term, the candidate term for the collocation can be identified. The collocation will be selected from the term that has a high frequency of co-occurrence with the ambiguous term. However, some words cannot be treated as collocation although they have a high frequency of co-occurrence with the ambiguous term. It is because collocation does not always occur in a fixed phrase and a high frequency co-occurrence term might be one of the function words, for example the word "di", "ialah", and "itu" in Malay. In order to remove all these function words, standard deviation of all co-occurrence terms will be calculated based on the range of the term located with the ambiguous term and the term with a low standard deviation will be selected [13]. The result of this stage is also essential for the next stage. In the next stage, all texts with a similar collocation will be grouped together with their respective sense. This step is done based on the one sense per collocation theory [19]. The result in this step is

shown in Table 3.

The following step is adapted from the knowledge-based approach; in this step, all selected texts with a similar collocation will be compared with the definition of the term in a dictionary in order to identify the sense which they represent in the dictionary. This step will be done by using the vector space model with the cosine similarity measure. The definition with the highest similarity value will be selected as the right sense and will be tagged into the text document. After that, all the tagged documents will be stored in a database and treated as the Malay tagged word sense corpus. This corpus is also applicable for word sense disambiguation using the supervised approach. Below is the similarity formula using the cosine similarity measure [20].

Table 3
Sense Group by Collocation

Term	Text	Collocation
daki	"Kadang-kadang kami daki bukit tetapi jumpa bukit yang sama juga," katanya.	bukit
daki	Dia daki bukit, panjat curam terjun curam dan sanggup berdepan	bukit
daki	Dia sampai tua asyik daki bukit saja meracik tekukur	bukit
daki	punggung yang tiga suku terdedah itu tebal diselaputi daki hitam yang bertompok-tompok macam lorek air ludah basi	hitam
daki	Toner Peluntur Daki - Tanggalkan Daki Hitam	hitam

$$Sim(\underset{d}{\rightarrow},\underset{q}{\rightarrow}) = \frac{\underset{d}{\rightarrow} . \underset{q}{\rightarrow}}{|| \underset{d}{\rightarrow} || \, || \underset{q}{\rightarrow} ||} = \frac{\vec{d} \cdot \vec{q}}{\sqrt{\sum_i d_i^2}\sqrt{\sum_i q_i^2}} = \text{Cosine}\,(\underset{d}{\rightarrow},\underset{q}{\rightarrow}) \quad (1)$$

where: $\underset{d}{\rightarrow}$ = Document or first vector
$\underset{q}{\rightarrow}$ = Query or second vector

The formula illustrates the similarity measure between two vectors which are vector d for the document and vector q for the query. The similarity will be calculated by identifying the cosine between these two vectors. The similarities will reach reliability and become more accurate upon reaching the nearest to one.

V. Conclusion

In conclusion, it can be seen that among the three primary methods in word sense disambiguation, the unsupervised method was the most selected by researchers to resolve the problem of ambiguity in a document and query. However, this method is said to have a slightly lower precision compared to the supervised method. This method has great potential as it is not limited, does not require human effort, and it is not hindered by the size of a lot of sense tagged corpus as happened in the supervised word sense disambiguation. This method manages to achieve high precision and is at times on par with the supervised approach. Therefore, the unsupervised and knowledge-based methods have been proposed for solving the problem of ambiguity in the Malay language document and query. This method is expected to provide a better result for Malay word sense disambiguation.

References

[1] R. Mihalcea, "Word sense disambiguation," *Encycl. Mach. Learn.*, pp. 1027–1030, 2010.

[2] O. Bakaya and D. Jurgens, "Semi-supervised learning with induced word senses for state of the art word sense disambiguation," *J. Artif. Intell. Res.*, vol. 55, pp. 1025–1058, 2016.

[3] K. Mittal and A. Jain, "Word sense disambiguation method using semantic similarity measures and owa operator," *CTACT J. Soft Comput. Spec. Issue Soft–computing Theory, Appl. Implic. Eng. Technol.*, vol. 5, no. 2, pp. 896–904, 2015.

[4] I. Lopez-Arevalo, V. J. Sosa-Sosa, F. Rojas-Lopez, and E. Tello-Leal, "Improving selection of synsets from WordNet for domain-specific word sense disambiguation," *Comput. Speech Lang.*, vol. 41, pp. 128–145, 2017.

[5] O. Ba and D. Jurgens, "Word sense induction and disambiguation rivaling supervised methods," vol. 1, no. 1993, pp. 1–15, 2013.

[6] A. Ranjan Pal and D. Saha, "Word sense disambiguation: a survey," *Int. J. Control Theory Comput. Model.*, vol. 5, no. 3, pp. 1–16, 2015.

[7] R. Navigli, K. C. Litkowski, and O. Hargraves, "SemEval-2007 Task 07: coarse-grained english all-words task," in *Proc. 4th Int. Work. Semant. Eval.*, 2007, pp. 30–35.

[8] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annu. Meet. Assoc. Comput. Linguist.*, 1995, pp. 189–196.

[9] A. Yusof, "Perluasan makna perkataan bahasa melayu: sumbangan data korpus berkomputer," pp. 1–14, 2009.

[10] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries," in *Proc. 5th Annu. Int. Conf. Syst. Doc. - SIGDOC '86*, 1986, pp. 24–26.

[11] S. S. Sazali, Z. Abu Bakar, and J. Jaafar, "Word prediction algorithm in resolving ambiguity in Malay text," pp. 1347–1352, 2016.

[12] B. Kapse and U. Shrawankar, "Word prediction using B+ tree for braille users," in *2013 Students Conference on Engineering and Systems, SCES 2013*, 2013.

[13] H. Manning, Christopher; Schütze, "Collocations," in *Foundations of Statistical Natural Language Processing*, 1999, pp. 151–189.

[14] M. Zakree, A. Nazri, S. M. Shamsudin, and A. A. Bakar, "An exploratory study of the Malay text processing tools in ontology learning," in *ISDA '08 Proc. 2008 Eighth Int. Conf. Intell. Syst. Des. Appl.*, 2008, pp. 375–380.

[15] F. Yahaya, N. A. Rahman, and Z. A. Bakar, "Resolving Malay word sense disambiguation utilizing cross-language learning sources approach conference," *Adv. Sci. Lett.*, vol. 4, no. 2, pp. 400–407, 2011.

[16] M. T. Pilehvar, D. Jurgens, and R. Navigli, "Align , disambiguate and walk: a unified approach for measuring semantic similarity," in *Proc. 51st Annu. Meet. Assoc. Comput. Linguist.*, 2013, pp. 1341–1351.

[17] D. Bouamor, G. Cedex, and P. Zweigenbaum, "Context vector disambiguation for bilingual lexicon extraction from comparable corpora," pp. 759–764, 2013.

[18] W. Gomaa and A. Fahmy, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[19] R. Navigli, "Word Sense Disambiguation: A Survey," *ACM Comput. Surv*, vol. 41, no. 10, 2009.

[20] S. Clark, "Vector space models of lexical meaning," *Handb. Contemp. Semant.*, no. March, pp. 1–43, 2014.