Detection of Compound Word with Combination Noun and Adjective using Rule Based Technique in Malay Standard Document

Zamri Abu Bakar, Normaly Kamal Ismail and Mohd Izani Mohamed Rawi
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia.
zamri@salam.uitm.edu.my

***Abstract*—In this paper we describe our methods for detecting the compound word with combination of Noun and Adjective Compound Nouns in Malay standard document. We addressed the problem on detection of combination noun and adjective in Malay sentences to become a compound word. We modified several identification rules based by using Malay grammar rules and syntactic information to increase the percentage of recall, precision and F1-Score. For compound word identification, we used dictionary-based and thesaurus information for implementing Part of Speech (POS) tagging to all words in the selected Malay document. Testing was done on selected Malay document. The result showed an improvement compared to previous research with a precision of 90.9%, a recall of 10.2% and a F1-Score of 18.1%.**

***Index Terms*—Compound Word; Malay Standard Document; Ruled-Based; Syntactic Information.**

I. INTRODUCTION

Research on compound words has been starting in English. Thus, researchers such as Veronika et al. [1] have discussed the fundamental concept of compound words in English. Most of English compound word is constructed by nouns and it was modified by other nouns, verbs and adjectives [2]. Similar concept was found in Malay compound word where by most of first words are noun and the following word is noun, verb, adjective or prepositional [3]. When it mentioned about noun, verb, adjective and prepositional, all of it we called as part of speech (POS) or in Malay called as "golongan kata".

The compound words are a word that is very productive and arguably day to day, the new compound words are created to describe the specific meaning of the language terms. So, it is not reasonable to store all the compound words in the dictionary. According to [4], since compounds are widely used in communication and writing, it is not impossible to magnify completely in the dictionary. If the compound words to be manually identified, it will jeopardize cost and time in order to add or update the dictionary. Thus, it is a necessity to automatically extract them into the dictionary before they will be translated to other languages [4]. Translations of this compound word should also be made so that the original meaning is similar when translated to other languages [5].

According to [6], recognition of compound noun in Malay sentence become one of the important thing because compound noun is commonly used in the following applications such as detecting the head and modifier of the words, extracting various knowledge from texts, analyzing the morphological, retrieving pattern documents and correcting grammatical structure of the phrase or sentence.

II. RELATED WORK

Grammar is a field which focused on word formation and process of making a sentence in any language [7]. According to Siti Hajar and Kamaruddin, grammar is a set of rules on how a certain word formed and how that word is combined with other words to produce grammatical sentence [8]. Robin mentioned that grammar is something that is compulsory in structure of speech's order or writing and a classification of repeated speech element based on its formulas [9].

According to Abdullah the compounding is a process of linking two or more basic words and carries certain meaning [33]. Compounding the words may be hyphenated, written open (as separate words), or written solid (closed) [34]. But this study will focus on compounding with separating between two words. An open compound is a combination of words with closely associated that they convey the idea of a single meaning [35]. Referring to [36], they mentioned that rules for performing the compounding of words are different in every language. In addition to that statement, discussion as in [37, 38] stated that research on compounding word is now very active in linguistic language and computational linguistic.

According to [39], the existing dictionary does not collect those new compound-words in time and does not correctly identify the word specifically. Therefore, they have presented the new method to solve the problem of compound-words in the field of information security such as semi-automated identification method [39]. The compound noun construction processes for Malay sentence characterize the words based on the combination of 1) noun and noun 2) noun and noun modifier 3) noun and non-noun modifier [6]. Veronika et al. [1] have used 20 types of noun modifier relationship to represent the semantic relations between concepts including agent, beneficiary, cause, instrument and etc. Therefore, the relationship types useful to get the right compound order for the words. CARIN model process involved the thematic relations, in which two steps must be implemented such as developing taxonomies of relations and identify and create the list of words [10]. Alfred et al. [11] have approached the name of dependency relations to identify the position of words located in compound noun as a head modifier or modifier head. They have found that the type relationships need to be used to analyze the input sentence in the structure of dependency triples, hierarchy of type dependencies and

syntactic level of words. However, not all the relations of recognizing the head modifier for Malay compound noun used in the explained structures in their research work. Muneer et al. [42] have presented the empirical result of sixteen statistical association measures of Malay <N+N> compound nouns extraction and the experimental results obtained are quite satisfactory in terms of the Precision, Recall and F-score.

Alfred et al. mentioned in their research that in the process of information extraction for Named-Entity Recognition (NER) is very crucial in identifying and locating entities such as person, location and organization. The method used in this research is Malay ruled-based for effectiveness retrieving proper noun from Malay article. Three rules were developed; 1) rules for identifying a person-entity 2) location rule and 3) organization rule with three major step from tokenization, part-of speech tagging and then classified under proper nouns category into the rules for location and person prepositions [11]. Veronika et al. [1] described methods to detect noun compounds and light verb construction in their test experiments. The three methods such as noun compounds, dictionary-based methods and POS-tagging contribute most in the performance of the system where they produce the best result. According to [12], the fundamental grammar rules determine grammatical behaviour such as the placement of word, verb agreement and passivity behaviour. This study focused on Arabic GR-related problems which pay attention on the difficulty of determining grammatical relations in Arabic sentences. Therefore, they have developed an effective fundamental grammar rules extraction technique for analysing Arabic standard sentences and ca out with an optimum solution [12]. Veronika et al. have proposed their technique for detecting the countability of English compound noun. The English compound nouns made up of two or more words and be formed by other nouns or objectives [1]. A simple detecting algorithm such as viable n-gram models where the parameters can be obtained by using WWW search engine such as Google. The output from algorithm proposed by Jing and Kenji could perform with 89.2% of 70.4% on the total test set. They classified the English compound noun into three classes such as countable, uncountable and plural. They obtained the information about countability of individual nouns easily from grammar books or dictionaries [13].

Motoki and Vitaly explained that the term in Japanese Language is known as the challenging problem in Natural Language Processing (NLP) because the nouns are combinations of single nouns and produce different meaning compared to basic nouns. Therefore, they have used a tool called TeamExtract to preserve text semantics by using online resources such as ALC online dictionary, Wikipedia and Google phrase service [14]. Jeena has discussed regarding the use of Sandhi rules for Malayalam compound words. The challenges of designing Sandhi rules generator as a standalone development system environment has been described in their research. He has studied and described the Sandhi Rules developed for four major Malayalam sentences [15]. Nair and David has developed and algorithm used for splitting the compound words and the splitter used for a full-fledged morphological analyser. The splitter has been developed to split a compound words into morphemes and the splitter used a lexicon tree that can reduce the loop times for the morphemes. An algorithm of splitter used is depth first strategy and can almost split all kinds of compound nouns in Malayalam [16].

Abdulgabbar and Juzaiddin have identified that the lexical units of compound noun is a very important task in NLP applications. Thus, they used hybrid method for extracting the noun compound from Arabic words based on linguistic knowledge and statistical measures [17]. Poria et al. [18] reviewed the proposed novel rule-based product in order to solve the problem of extraction which exploits knowledge and sentences dependency trees to detect both explicit and implicit aspects. They have reviewed the two popular dataset in evaluating the system through an extraction technique with obtained the higher detection accuracy for both datasets. Lishu et al. [19] have created a novel neural network to stimulate the recognition process of compound in English and Chinese. Rule based approach is still used in processing natural language because its rule relies in solid linguistic knowledge. Although many approaches have been proposed, an automatic extraction of compound word is still a major area of research primarily because the effectiveness of current automated compound noun extraction is not produce more desirable result and still needs improvement in terms of dependency relationships [11]. There are various methods to extract compound word from the corpus and it was proposed by some researchers. Among them is to use a statistical method such as mutual information in which these methods can be used with norm-based association and dependency of the context [20, 21]. Gao et al. obtained the candidate list of compound word from the corpus by using concept of entropy from information theory such as decision tree learning [22]. Meanwhile, Luo et al. [23] also proposed the statistical approach to extract words by checking the characters in word itself. However, Xiong and Zhu used a method which combined relative frequency between mutual information and entropy that changes frequency arrangement from ordered into disordered in isolated systems. According to research works above, it can be inferred that they used statistical information to process the texts from the corpus [24]. Mutual Information [MI] is a standard measure of the strength of association between co-occurring items and has been used successfully in extracting collocations from English [25] and performing Chinese word segmentation [21,26-28]). According Alfred et al. [11], the effectiveness of extraction compound nouns based on measurement of recall and precision, the result showed that precision is given with better output but in terms of recall, the result is quite worse.

Motoki and Vitaly explained that they can only detect 26.5% compound noun using TermExact. This TermExact use the method based on Occurrence and Concatenation Frequency [14]. Suhaimi et al. [41] have extracted the compound word using the Subject Verb Modifier (SVM) method and statistical co-occurrence. This method is referred to rule-based method. Based on the result from the experiment, they only achieved a precision of 82.25% and recall of 77.44%. This study use data from the website and they specific the type of data limited to the information security. The data collection consists of 20,000 pieces of news from www.sina.com.cn. Goncalves et al. [29] conducted experiments to extract the multi word using parallel algorithm. They used corpus up to 1 billion words from testing data.

Maryam and Nazlia [30] have conducted the extraction of nested noun compound for Arabic Language using the hybrid method between linguistic approach and statistical method. This study tried to obtain the nested compound noun such as the multiword expression “enterprise resource planning” can

have two compound nouns which are enterprise resource and resource planning. To get this compound noun, the n-gram method also has been done to cater all the sentences.

III. RESEARCH METHODOLOGY

Four phases were identified in the proposed method as in Figure 1 have been used in this study. The phases include: (i) corpus acquisition for the input: (ii) pre-processing tasks that consist of three tasks which is normalization, stemming and tokenization: (iii) the extraction of the compound noun extraction consists of POS tagging, thematic relation detector and head modifier generator: and (vi) modified Malay grammar rules for compound nouns extraction method. Malay grammar rules have been modified in the database to improve the performance of the method. With the modification of the rules, it has improved the effectiveness of the extraction of compound nouns.

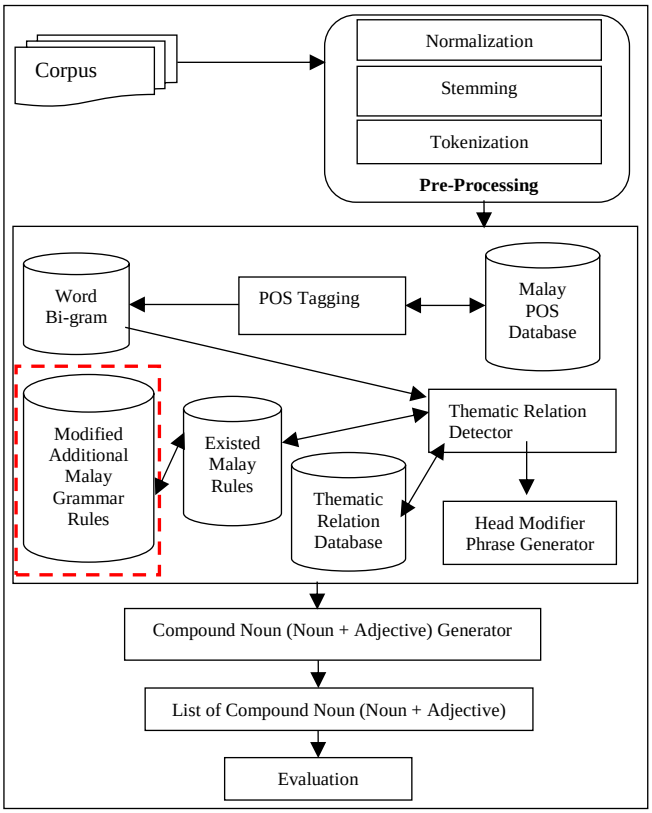

Figure 1: Research Design Framework of the Proposed Method

A. Corpus Acquisition

In this step, this study will collect all types of Malay article such as website blog, official website, story books, dictionary, school text books, magazines, newspapers and a sample student assay for UPSR, PT3 and SPM. This study estimated to have 3,124 sentences to be processed and will extract the compound noun from the Malay news which is Utusan Malaysia.

Five phases identified the proposed method that has been used in this study. The phases include; (i) corpus acquisition for the input, (ii) pre-processing tasks that consists of three tasks which are normalization, stemming and tokenization (ii) the extraction of the compound noun extraction that consists of POS tagging, thematic relation detector and head modifier generator: and (vi) the evaluation metric, this phased are used to evaluate the method that has been proposed.

Malay grammar rules will be added in the database to improve the performance of the method. With the modification of the rules, it can be expected to improve the effectiveness of the extraction of compound nouns. Candidate ranking aims to determine the association measures for the extracted candidates in the bi-gram lists where it allocates to each candidate a score of association strength.

B. Pre-Processing

In this phase, the Malay word is the first process of this part. Below is an example of tagging process for the selected Malay word. This process is done manually by referring to [40]. In this phase, based on pilot study, this study choosed to use the newspaper corpus as a training data to be processed, while all crawled websites are proposed by removing HTML tags, identifying main content, automatic noise removal and breaking the content down to sequence of individual tokens. After that, all-uppercase, capitalizes and mixed case words were changed to lowercase format. Punctuations, special symbols and numbers were removed. Tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation.

C. Compound Word Candidate Generation

In this stage, this study has tagged all nouns, verbs and adjectives word in the text corpus. This phase gave all possible noun-verb collocations that occur in a corpus [16]. From the tagging process of corpus, if consecutive words tagged as Noun and Verb has been extracted as a candidate for the compound nouns. These compound noun candidates are passed to the next phrase for automatic compound nouns extraction method. Other than that, this study goes through processing to identify the compound noun candidate that occurs with the frequency in the corpus which are greater than or equal to two. For this stage, the method that was applied by Ong has been used as it compatible and has been applied for the extraction of compound noun in the Malay Language [31]. However, the seven Malay Grammar Rules which was used by [31] have been modified to include additional Malay grammatical rules to improve the possible number of candidates obtained [6].

D. Pre-Processing Modified Grammar Rules for Compound Word Extraction Method

When we extract the compound words from the compound word candidate generation phase, this study proposes Malay grammar rules to detect the compound word candidate from the Malay corpus. Thus, it means that, this study proposes linguistic knowledge approach to extract and classify the compound nouns from the Malay corpus [31]. In order to extract compound nouns in standard Malay sentences, the first step is, to understand the Malay grammar. Basically, Malay grammar explained that the sentence must have a subject, verb and predicate [32]. In this study, the sentence has been chucked into several simple sentences from a long sentence. For example:

<u>Sentence 1:</u>

Saban tahun, jumlah pelancong **yang** direkodkan **semakin** bertambah **bukan sahaja** pelancong **dari** negara jiran **seperti** Singapura, Thailand **dan** Indonesia **malah** turut **menjadi** tumpuan pelancong asing **dari** Asia Barat, Eropah **dan** Amerika Syarikat.

The first step is, we removed all auxiliary word (*kata bantu*), conjunction word (*kata hubung*), *kata sendi* and *kata*

pemeri which are *adalah, ialah, yang, semakin, bukan, sahaja, dari, seperti, dan* and *malah*. Besides that, we also removed the comma. Then, the sentence becomes a simple phases. Below is a phrase sentences after the removal process has been done.

Saban tahun || jumlah pelancong || direkodkan || bertambah || pelancong || negara jiran || Singapura || Thailand || Indonesia || turut || tumpuan pelancong asing || Eropah || Amerika Syarikat ||

The next step is Part of Speech (POS) tagging where by all the word in the sentence are tagged automatically. Below is an example of tagging process to each of the word.

Saban[KN] tahun[KN] || jumlah[KN] pelancong [KN] || direkodkan [KK] || bertambah [KK] || Pelancong [KN] || negara [KN] jiran[KN] || Singapura [KN] || Thailand [KN] || Indonesia [KN] || urut[KK] || tumpuan[KN] pelancong[KN] asing[KN][KK] || Eropah[KN] || Amerika[KN] Syarikat[KN] ||.

From this phrase, we chose the combination of two words with co-occurrence of both words which are Noun (*Kata Nama [KN]*) and Adjective (*Kata Adjektif [KA]).* At the same time, we have also removed all combination word with *Kata Nama [KN]* and Determiner (*Kata Penentu [KP]*) which was "*itu", "ini", "tersebut"* and etc.

IV. Result and Evaluations

To organize our results and evaluations, we divided them into two different groups of data samples. The first is in the training data set which contains 3,124 samples of the Malay sentences, while 765 samples of the Malay sentences are in the testing data set. In Figure 2 below, it shows a few examples of compound noun that was taken out from the testing sentences using the summation algorithm in the previous discussion part.

The comparison of the results is shown in Table 1. Finally, the recall, precision and F-Measure value by using modified Malay grammar Rules is increased to 0.2 percent. The percentage of improvement is slightly lower but it is significant for this study. This study has assisted for increasing the percentage values of improvement result. The Performance and Accuracy Measurement are described below;

$$\text{Precision} = \frac{X}{(X+Z)} \times 100 \quad (1)$$

$$\text{Recall} = \frac{X}{(X+Y)} \times 100 \quad (2)$$

where: X = The total compound word retrieved to the query
Y = The total compound word not retrieved that relevant to the query
Z = The total not relevant compound word retrieved

So, the precision and recall are evaluated as follows:
X = Total relations relevant
Y = (Number of records on a particular topic – Total relations relevant)
Z = (Total relations retrieved - Total relations relevant)

The measurement of the harmonic mean for recall and precision is formulated using this F1-score equation:

$$\text{F1-score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (3)$$

adat hidup	anak jati	badan kasar
adat resam	anak kedua	baginda mangkat
agama bersepadu	anak sulung	bahagian lain
agama sempurna	anak yatim	bahagian utama
agama terus	ancaman lain	bahan cemar
agensi berkaitan	anggota tetap	baju biru
agensi lain	anugerah pertama	bandar baru
air besar	api kecil	bangunan baharu
akhbar pertama	aplikasi berkaitan	bangunan bersejarah
aksi akhir	aplikasi terbesar	bangunan utama
aktiviti awal	aras tinggi	bank tempatan
aktiviti lasak	artis baharu	bantuan segera
aktiviti luar	artis terkenal	barang kemas
aktiviti sama	askar wataniah	barangan mewah
aktiviti wajib	aspek jasmani	barangan segar
alam fana	aspek teknikal	batalion pertama
alam maya	aspek utama	batang utama
amaran awal	atlet lain	baucar bernilai
amaran keras	badan kasar	amaran keras
alam maya	bandar baru	badan halus
amaran awal		

anggota tetap *(permanant employee)*
api kecil *(small fire)*
aktiviti lasak *(extream activity)*
bandar baru *(new town)*
bangunan bersejarah *(historical building)*
baju biru *(blue shirt)*
bank tempatan *(local bank)*

Figure 2: Example of Compound Words

Table 1
Recall, Precision and F1-Score for Each Relation

Word Combination	Baseline			Modified Malay Grammar Rules		
	Precision (%)	Recall (%)	F-Measure (%)	Precision (%)	Recall (%)	F-Measure (%)
Noun & Adjective	90.4	9.7	17.6	90.9	10.2	18.1

Finally, after finishing the experiment process, the system listed out all compound words. All the compound words are validated by expert to prove that all the listed compound nouns are correct based on the Malay grammar definition.

Rules in a Malay noun phrase can be recognized using a dependency relationship approach. The result shows significant improvement in terms of the effectiveness for the relationship types used. This is done by evaluating them with the baseline values compiled from a set of training and testing data from our study. However, the percentage value of evaluation by using precision, recall and F1-Sore measurement is produced with not slightly higher but it showed the improvement of the modification of the technique. This is because of the lack of test data required in our testing process. In future research work, we will improvise the structure of Malay sentence to become an additional part of Malay grammar rules structure. The use of larger data is also required in the training and test dataset for the experiment to get better results.

ACKNOWLEDGMENT

Authors would like to thank Utusan Malaysia Bhd for the test sampling data on Malay text document. Authors also would like to Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia for the support grant.

REFERENCES

[1] V. Vincze, T. I. Nagy, and G. Berend , "Detecting noun compounds and light verb constructions: A contrastive study," in *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World (MWE 2011)*, 2011, pp. 116-121.

[2] V. Nastase and S. Szpakowicz, "Exploring noun-modifier semantic relations," in *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-03)*, 2003, pp. 285-301.

[3] S. A. Rahman, N. Omar, and N. B. C. Hassan, "Construction of compound nouns (CNs) for noun phrase in Malay sentence," in *Proceedings of the 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, 2012, pp. 22–25.

[4] K. Y. Su, M. W. Wu, and J. S. Chang, "A Corpus-based Approach to Automatic Compound Extraction," in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, 1992, pp. 242-247.

[5] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword expressions: A pain in the neck for NLP," in *Computational Linguistics and Intelligent Text Processing*, 2002, pp. 38-43.

[6] S. A. Rahman, N. Omar, and J. A. Aziz, "Extraction of compound nouns in Malay noun phrases using a noun phrase frame structure," *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 3, no. 1, pp. 23-32, 2014.

[7] F. M. M. Sultan, "Struktur sintaksis frasa nama (FN) Bahasa Melayu," *Jurnal Bahasa*, vol. 8, pp. 204-219, 2008.

[8] S. H. A. Aziz and K. Husin, *Pusat Sumber Sekolah"*. Kuala Lumpur: Kumpulan Budiman, 1996.

[9] R. H. Robin, *General Linguistics. An Introductory Survey*. London : Longman, 1971.

[10] S. A. Rahman, N. Omar, and M. J. A. Aziz, "A fundamental study on detecting head modifier noun phrases in Malay sentence," in *Proceedings of the International Conference on Semantic Technology and Information Retrieval*, *STAIR*, 2011, pp. 255-259.

[11] R. Alfred, L. C. Leong, C. K. On, and Anthony, P., "Malay named entity recognition based on rule-based approach," *International Journal of Machine Learning and Computing*, vol. 4, no. 3, pp. 300-306, 2014.

[12] M. A. Falih and N. Omar, "A comparative study on Arabic grammatical relation extraction based on machine learning classification," *Middle-East Journal of Scientific Research*, vol. 23, pp. 1222-1227, 2015.

[13] J. Peng and K. Araki, "Detecting the countability of English compound nouns using web-based models," *International Joint Conference on Natural Language Processing*, 2005, pp. 103-107.

[14] M. Miyashita and V. Klyuev, "TermExtract: Accuracy of compound noun detection in Japanese," in *Future Information Technology*, J. J. Park, Y. Pan, C.-S. Kim, and Y. Yang, Eds. Berlin, Heidelberg: Springer, 2014, vol. 276, pp. 473-476.

[15] J. Kleenankandy, "Implementation of Sandhi-rule based compound word generator for Malayalam," in *Proceedings of the Fourth International Conference on Advances in Computing and Communications*, 2014, pp. 134-137.

[16] L. R. Nair and S. D. Peter, "Development of a rule based learning system for splitting compound words in Malayalam language," *IEEE Recent Advances in Intelligent Computational Systems*, 2011, pp. 751-755.

[17] A. M. Saif and M. J. A. Aziz. "An automatic noun compound extraction from Arabic corpus," in *Proceedings of the International Conference on Semantic Technology and Information Retrieval, STAIR 2011*, 2011, pp. 224-230.

[18] S. Poria, E. Cambria, L. W. Ku, C. Gui, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," in *Workshop on Natural Language Processing for Social Media (SocialNLP)*, 2014, pp. 28-37.

[19] L. Li, J. Chen, Q. Chen, and F. Fang, "A novel model for recognition of compounding nouns in English and Chinese," in *Proceedings of the 6th International Symposium on Neural Networks, Part III ISNN 2009 Wuhan, China*, 2009, pp. 457-465.

[20] L.F. Chien, "PAT-tree-based keyword extraction for Chinese information retrieval," in *Proceedings of the ACM SIGIR'97 Conference*, 1997, pp. 50-58.

[21] J. Zhang, J. Gao, and M. Zhou, "Extraction of Chinese compound words - An experimental study on a very large corpus," in *Proceedings of the Second Workshop on Chinese Language Processing Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, vol. 12, pp.132.

[22] J. F. Gao, J. Goodman, M. J. Li, and K.F. Lee, "Toward a unified approach to statistical language modeling for Chinese," *ACM Transactions on Asian Language Information Processing*. Vol. 1, no. 1, pp. 3-33, 2002.

[23] Luo, S. F. and Sun, M. S., "Two-character Chinese word extraction based on hybrid of internal and contextual measures," in *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 2003, 2003, pp. 24-30.

[24] Y. Xiong J. and Zhu, "Toward a unified approach to lexicon optimization and perplexity minimization for Chinese language modeling," in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, 2005, pp. 18-21.

[25] D. Lin, "Automatic retrieval and clustering of similar words," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics ACL'98 and 17th International Conference on Computational Linguistics*, 1998, pp.768-774.

[26] R. Sproat and C. Shih, "A statistical method for finding word boundaries in Chinese text," *Computer Processing of Chinese & Oriental Languages*, vol. 4, pp. 336-351, 1990.

[27] S. Maosong, S. Dayang, and B. K. Tsou, "Chinese word segmentation without using lexicon and hand-crafted training data," in *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics*, 1998, pp. 1265-1271.

[28] S. Peng and D. Schuurmans, "Self-supervised Chinese word segmentation," in *Advances in Intelligent Data Analysis*, F. Hoffmann, D. J. Hand, N. Adams, D. Fisher, and G. Guimaraes, Eds. Berlin, Heidelberg: Springer, 2001, pp. 238-247.

[29] C. Goncalves, J. F. Silva, and J. C. Cunha, "A parallel algorithm for statistical multiword term extraction from very large corpora," in *Proceedings of the IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security and 2015 IEEE 12th International Conference on Embedded Software and Systems*, 2015, pp. 219-224.

[30] M. Al-Mashhadani and N. Omar, "Extraction of Arabic nested noun compounds based on a hybrid method of linguistic approach and statistical methods," *Journal of Theoretical and Applied Information Technology*, vol. 76, no. 3, pp. 408-416, 2015.

[31] O. C. Guan, *Kuasai Struktur Ayat Bahasa Melayu.* Malaysia: Dewan Bahasa dan Pusataka (DBP), 2009.

[32] A. Hassan, *Linguistik Am*. Malaysia: PTS Profesional Publishing, 1992.

[33] A. Hassan, *Tatabahasa Bahasa Melayu: Morfologi dan Sintaksis*. Malaysia: PTS Publications and Distributors, 2002.

[34] A. K. M. Nor, *Tatabahasa Asas*. Kuala Lumpur: Persatuan Pendidikan Bahasa Malaysia, 2012.

[35] B. J. Juhasz, Y. H. Lai, and M. L. Woodcock, "A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience," *Behavior Research Methods*, vol. 47, no. 4, pp. 1004-1019, 2015.

[36] J. Palemans, K. Demuynck, H. V. Hamme, and P. WamBacq, "Coping With Language Data Sparsity: Semantic Head Mapping of Compound Words," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014, vol. 6, pp. 1-5.

[37] D. ´O S´eaghdha, *Learning Compound Noun Semantics*. Ph.D. Thesis, Computer Laboratory, University of Cambridge, 2008.

[38] I. Hendrickx, Z. Kozareva, P. Nakov, D. ´O S´eaghdha, S. Szpakowicz, and T. Veale, "Semeval-2013 task 4: Free paraphrases of noun compounds," in *Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 138-143.

[39] S. Li, L. Zhang, B. Han, T. Lei, Q. Wang, T. Peng, and P. Cao, "A SVM-based compound-word recognition method in information security," in *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2013, pp. 837-841.

[40] M. Taharim, R. Ja'afar, and N. A. Shukur, *Tesaurus Bahasa Melayu Dewan*. Malaysia: Dewan Bahasa dan Pustaka (DBP), Edisi Baharu, 2015.

[41] S. A. Rahman, N. Omar, and M. J. A. Aziz, "The effectiveness of using the dependency relations approach in recognizing the head modifier for malay compound nouns," in *2014 International Conference Computer and Information Sciences (ICCOINS)*, 2014, pp. 837-841.

[42] M. A. S. Hazaa, N. Omar, F. M. Ba-Alwi, and M. Albared, "Automatic extraction of Malay compound nouns using a hybrid of statistical and machine learning methods," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 3, 925-935, 2016.