Rule-based Approach on Extraction of Malay Compound Nouns in Standard Malay Document

Zamri Abu Bakar, Normaly Kamal Ismail, Mohd Izani Mohamed Rawi
Faculty Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, 40450, Malaysia

Corresponding author: zamri@salam.uitm.edu.my, normaly@tmsk.uitm.edu.my, izani@tmsk.uitm.edu.my

Abstract. Malay compound noun is defined as a form of words that exists when two or more words are combined into a single syntax and it gives a specific meaning. Compound noun acts as one unit and it is spelled separately unless an established compound noun is written closely from two words. The basic characteristics of compound noun can be seen in the Malay sentences which are the frequency of that word in the text itself. Thus, this extraction of compound nouns is significant for the following research which is text summarization, grammar checker, sentiments analysis, machine translation and word categorization. There are many research efforts that have been proposed in extracting Malay compound noun using linguistic approaches. Most of the existing methods were done on the extraction of bi-gram noun+noun compound. However, the result still produces some problems as to give a better result. This paper explores a linguistic method for extracting compound Noun from stand Malay corpus. A standard dataset are used to provide a common platform for evaluating research on the recognition of compound Nouns in Malay sentences. Therefore, an improvement for the effectiveness of the compound noun extraction is needed because the result can be compromised. Thus, this study proposed a modification of linguistic approach in order to enhance the extraction of compound nouns processing. Several pre-processing steps are involved including normalization, tokenization and tagging. The first step that uses the linguistic approach in this study is Part-of-Speech (POS) tagging. Finally, we describe several rules-based and modify the rules to get the most relevant relation between the first word and the second word in order to assist us in solving of the problems. The effectiveness of the relations used in our study can be measured using recall, precision and F1-score techniques. The comparison of the baseline values is very essential because it can provide whether there has been an improvement in the result.

1. Introduction
There are many natural languages available throughout the world. Natural language is a language which has been used by human beings in daily life for communicating, presenting or conveying information and in knowledge learning process. According to [1] to date, there are about 6912 speaking languages which are now being used by human beings worldwide. However, Mandarin, English, Hindi, Spanish, Arabic, Russian, Malay, Portuguese, Bengali and French are the ten biggest speaking languages with regard to the total number of speakers. The Malay language is ranked at 7th
with approximately 259 million speakers. Among those biggest speaking languages, the Malay language will be chosen as the main topic or focus in this study. It is because according to [2], research in the Malay language document is still underdone and too little among researchers in Malaysia, Dewan Bahasa dan Pustaka (DBP) is established as a government body to coordinate and empower the Malay language as the national language and official language of the country. The researches that have been conducted by [2] - [4], discussed on the Malay grammar to strengthen the structure of sentences and this will reinforce the usage of the Malay language.

In natural language text, the compound noun is a word that is very productive and arguably day to day, and the new compound word is 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 [5], since compounds are widely used in communication and writing, it is not impossible to completely magnify then in the dictionary. If the compound words are 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. Translations of this compound word should also be made so that the original meaning is similar when translated to other languages [6].

According to [7], recognition of compound noun in Malay sentences has become one of the important things 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.

2. Related Work

Grammar is a field which focused on word formation and process of making a sentence in any language [6]. According to [8], grammar is a set of rules on how a certain word is formed and how that word is combined with other words to produce grammatical sentence. According to [9], a linguist defines grammar as a group of system that must be obeyed by a language user and it becomes a basic concept in producing a beautiful language. [10] mentioned that grammar is something that is compulsory in the structure of speech order or writing and a classification of repeated speech element is based on its formula.

According to [11], rules for performing the compounding of words are different in every language. In addition to that statement, [12],[13] stated that a research on compounding word is now very active in linguistic language and computational linguistic. Referring to [14], the existing dictionary does not collect those new compound-words in time and does not correctly identify the word specifically. Therefore, they have presented a new method for solving the problem of compound-words in the field of information security such as semi-automated identification method. The compound noun construction process for Malay sentences characterizes the words based on the combination of; 1) noun and noun 2) noun and noun modifier: and 3) noun and non-noun modifier [14]. [9] has used 20 types of noun modifier relationship to represent the semantic relations between concepts including agent, beneficiary, cause, instrument and etc. The relationship types are useful to get the right compound order for the words. CARIN model process involves the thematic relations, in which two steps must be implemented such as developing taxonomies of relations, and identifying and creating the list of words [8]. Referring to [15] 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 analyse 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 are used in the explained structures in their research work. [16] 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.

[15] in their research stated that 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 steps from tokenization, part-of speech tagging and classified under proper nouns category into the rules for location and person prepositions. Referring to [17], they describe the methods to detect noun compounds and light verb construction in their test experiments. The three methods which are noun compounds, dictionary-based methods and POS-tagging contributed the most in the performance of the system where it produced the best result. According to [18], the fundamental grammar rules determine grammatical behaviour such as the placement of word, verb agreement and passivity behaviour. The study focused on Arabic GR-related problems in which they pay intention 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 have come out with an optimum solution. [17] have proposed their technique for detecting the countability of English compound noun. The English compound nouns are made up of two or more words and are formed by other nouns or objectives. The detecting algorithm is based on simple model such as viable n-gram model where the parameters can be obtained by using WWW search engine such as Google. The output from algorithm proposed by [19] could perform with 89.2% on the total test set. They classified the English compound noun into three classes which were countable, uncountable and plural. They obtained the information about countability of individual nouns easily from grammar books or dictionaries.

According to [20], the term in the 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. [21] has discussed regarding to the use of sandhi rules for Malayalam compound words. The challenging design of Sandhi rules generator as a standalone development system environment has been described in their research. [21] have studied and described the Sandhi Rules developed for four major Malayalam sentences. [23] has developed an algorithm used for splitting the compound words and the splitter is used for a full-fledged morphological analyser. The splitter has been developed to split a compound word into morphemes and the splitter used a lexicon tree that can reduce the loop times for the morphemes. An algorithm of splitter use is depth first strategy and almost splits all kinds of compound nouns in Malayalam.

According to [24], they 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. [26] reviewed the proposing of a novel rule-based product in order to solve the problem of extraction with exploited knowledge and sentences dependency trees to detect both explicit and implicit aspects. They reviewed the two popular dataset in evaluating the system through an extraction technique with obtained the a higher detection accuracy for both datasets. [26] have created a novel neural network to stimulate the recognition process of compound nouns in English and Chinese. Rule based approach is still being used in processing natural language because its rule relies on solid linguistic knowledge. Although many approaches have been proposed, automated automatic extraction of compound word is still a major area of research primarily because the effectiveness of current automated compound noun extraction does not obtain a better result and still needs improvement in terms of dependency relationships [15]. 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 this method can be used with norm-based association and dependency of the context [27],[28]. [30] obtained the candidate list of compound word from the corpus by using concept of entropy from information theory such as decision tree learning. Meanwhile, [31] also proposed the statistical approach to extract words by checking the characters in the word itself. However, [32] used combination of relative frequency between mutual information and entropy that nature tends from order to disorder in isolated systems. According to research works above, it can be inferred that they
used statistical information to process the texts from the corpus. Mutual Information is a standard measure of the strength of association between co-occurring items and has been used successfully in extracting collocations from English [33] and performing Chinese word segmentation [29],[34]-[36]).

According to [15], the Effectiveness of Extraction Compound Nouns based on measurement of Recall and Precision, result showed that Precision is given with better output but in terms of recall, the result is quite bad.

The research by [16] focused on the automatic extraction of the N-N Malay compound nouns multiword expression (MWE). Experiments presented by [16] were performed on Malay data and the attention was restricted to the first and second categories of Malay noun compounds. Refer to research paper [8].

[21] 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” that have two compound words which are enterprise resource and resource planning. To get this compound word, the n-gram method must also be done to cater this sentence.

3. Research Methodology

Four phases identified in the proposed method 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 word 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 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 words.

3.1. 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 word 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 word 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.

3.2. 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 [51]. In this phase, based on pilot study, this study will choose 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 are 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.

3.3. Compound Noun Candidate Generation
In this stage, this study will tag all nouns, verbs and adjectives word in the text corpus. This phase gives all possible noun-noun collocations that occur in a corpus [36]. From the tagging process of corpus, if consecutive words tagged as Noun and Noun 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 [7] will be used as it compatible and have been applied for the extraction of compound word in the Malay Language. However, the seven Malay Grammar Rules which was used by [7] will be modified to include additional Malay grammatical rules to improve the possible number of candidates obtained.

3.4. Modified Malay Grammar Rules for Compound Nouns 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 [7]. 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 [3]. In this study, the sentence has been chucked into several simple sentences from a long sentence. For example:

**Sentence 1:**
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 pemer 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 phrases. 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] || turut[KK] || tumpuan[KN] pelancong[KN] asing[KK][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 Kata Nama[KN] and kata Nama[KN]. At the same time, we also removed all combination words with KN and Kata Penentu(KP) which were itu, ini, tersebut and etc.

4. Results and Evaluations and Result
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 Table 1 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.
TABLE 1. Example of compound nouns

| Word Combination | Baseline | Modifier Malay Grammar Rules |
|------------------|----------|-----------------------------|
|                  | Precision (%) | Recall (%) | F-Measure (%) | Precision (%) | Recall (%) | F-Measure (%) |
| Noun Noun        | 96.9      | 55.3         | 70.4           | 97.1          | 55.5        | 70.7           |

The comparison of the results is shown in Table II. Finally, the recall and precision 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 will assist to increase the percentage values of improvement defined in our research objective.

TABLE II. RECALL AND PRECISION FOR EACH RELATION

5. Conclusion

We have discussed how the modified Malay Grammar 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 produced is not slightly higher due to 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.
6. Acknowledgement
We would like to thank the Development of Student and Academic Fund (TAPA), Faculty Computer and Mathematical Sciences, for supporting registration fees.

References
[1] Ethnologue, “Languages of the World”. (2012, August 28). Retrieved November 01, 2016, from http://www.ethnologue.com/
[2] Zuraidah Mohd Don, “ Processing Natural Malay Texts: A Data-Driven Approach”, 14, 90–103, 2010.
[3] Nik Saffiah Karim, Farid M Onn, & Hashim Haji Musa, “Tatabahasa Melayu”. Kuala Lumpur: Dewan Bahasa dan Pustaka, 2011
[4] Ong Ching Guan, “Kuasai Struktur Ayat Bahasa Melayu”, Dewan Bahasa dan Pustaka (DBP), Malaysia, 2009.
[5] Keh Yih Su, Ming Wen Wu, & Jing Shin Chang, “A Corpus-based Approach to Automatic Compound Extraction”, 1992.
[6] Sag Ivan A., Baldwin Timothy, Bond Francis, Copestake Ann, & Flickinger Dan, “Multiword expressions: A pain in the neck for NLP”. Computational Linguistics and Intelligent Text Processing, pp. 38-43, 2002
[7] Suhaimi Ab Rahman & Omar, N., “Transforming Noun Phrase Structure Form Into Rules To Detect Compound Nouns In Malay Sentences”. Journal of ICT, 12, 2013, Pp: 161–173, 161–17, 2013.
[8] Siti Hajar Hj abdul Aziz and Kamaruddin Hj Hussein, “Pusat Sumber Sekolah”. KL: Kumpulan Budiman.Sdn.Bhd, 1996.
[9] Chomsky, N. (2006), “Language and Mind Third Edition”. Massachusetts Institute of Technology, Cambridge University Press, 2006.
[10] Robin, R. H., “General Linguistics. An Introductory Survey”. London : Longman, 1971.
[11] Palemans. J., Demuynck, K., Hamme, H. V., & WamBacq. P., “Coping With Language Data Sparsity: Semantic Head Mapping of Compound Words”. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 6(1), 1–5, 2014.
[12] Diarmuid ’O S’eaghdha, “Learning compound noun semantics”. Ph.D. thesis, Computer Laboratory, University of Cambridge, 2008.
[13] Hendrickx, I., Kozareva, Z., Nakov,P., Diarmuid ’O S’eaghdha, Stan Szpakowicz, & Tony Veale, “Semeval-2013 task 4: Free paraphrases of noun compounds”. In Workshop on Semantic Evaluation (SemEval 2013), pages 138–143, 2013.
[14] Shixian Li, Lei Zhang, Bo Han, Tingrui Lei, Qing Wang, Tao Peng & Peng Cao, “A SVM-based Compound-word Recognition Method in Information Security”, 837–841, 2013.
[15] Alfred, R., Leong, L. C., On, C. K., & Anthony, P., “ Malay Named Entity Recognition Based on Rule-Based Approach”. International Journal of Machine Learning and Computing, 4(3), 300–306, 2014.
[16] Muneer A. S. Hazaa, Nazlia Omar, Fadl Mutaher Ba-Alwi, & Mohammed Albared, “Automatic Extraction Of Malay Compound Nouns Using A Hybrid Of Statistical And Machine Learning Methods”. International Journal of Electrical and Computer Engineering (IJECE), 6(3), 925, 2016.
[17] Veronika Vineze, Istvan Nagy T. & Gabor Berend, “Detecting noun compounds and light verb constructions: a contrastive study”. Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World (MWE 2011), Pages 116–121, (June), 116–121, 2011.
[18] Falih, M. A., & Omar, N., “A Comparative Study on Arabic Grammatical Relation Extraction Based on Machine Learning Classification”, 23(6), 1222–1227, 2015.
[19] Jing Peng & Kenji Araki, “Detecting the Countability of English Compound Nouns Using Web-based Models”, 103–107, 2005.
[20] Motoki Miyashita & Vitaly Klyuev, “TermExtract: Accuracy of Compound Noun Detection in Japanese”. Lecture Notes in Electrical Engineering, 276(1), 473–476. doi:10.1007/978-3-642-40861-8, 2014.

[21] Jeena Kleenankandy, “Implementation of Sandhi-Rule Based Compound Word Generator for Malayalam”. 2014 Fourth International Conference on Advances in Computing and Communications, 134–137, 2014.

[22] Maryam Al-Mashadani & Nazlia 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, 76(3), 408–416, 2015.

[23] Nair, L. R., & David Peter, S., “Development of a rule based learning system for splitting compound words in Malayalam language”. 2011 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2011, 751–755, 2011.

[24] Abdulgabar Mohammed Saif & Mohd Juzaidin Ab Aziz. “An Automatic Noun Compound Extraction from Arabic Corpus”. 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011, (June), 224–230, 2011.

[25] Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A., “A Rule-Based Approach to Aspect Extraction from Product Reviews”. Workshop on Natural Language Processing for Social Media (SocialNLP), 28–37, 2014.

[26] Lishu Li, Jiawei Chen, Qinghua Chen & Fukang Fang, “A Novel Model for Recognition of Compounding Nouns in English and Chinese”. 6th International Symposium on Neural Networks, ISNN 2009 Wuhan, China, May 26-29, 2009 Proceedings, Part III, 457–465, 2009.

[27] Chien, L. F., “PAT-tree-based keyword extraction for Chinese information retrieval”. In Proceedings of the ACM SIGIR '97 Conference, pp. 50-58, 1997.

[28] Zhang, J., Gao, J., & Zhou, M., “Extraction of Chinese compound words - An Experimental Study on a Very Large Corpus”. Proceedings of the Second Workshop on Chinese Language Processing Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics -, 12, 132, 2000.

[29] Gao, J. F., Goodman,J., Li, M. J. & Lee, K. F., “Toward a unified approach to statistical language modeling for Chinese”, ACM Transactions on Asian Language Information Processing, pp. 3-33, 2002.

[30] 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, pp. 24-30, 2003.

[31] Xiong, Y. & Zhu, J., “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, pp. 18-21, 2005.

[32] Lin, D., “Automatic retrieval and clustering of similar words”. ACL ’98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 768–774, 1998.

[33] Sproat, R. and C. Shih, “A statistical method for finding word boundaries in Chinese text”. Computer Processing of Chinese & Oriental Languages, 4, 4, 1990.

[34] Maosong, S., Dayang, S., & Tsou, B. K., “Chinese word segmentation without using lexicon and hand-crafted training data”. Proceedings of the 36th Annual Meeting on Association for Computational Linguistics -, 1265–1271, 1998.

[35] Peng, S. & Schuurmans, D., “Self-supervised Chinese Word Segmentation”. IDA 2001, LNCS 2189, pp. 238-247, 2001. Springer-Verlag Berlin Heidelberg, 2001.

[36] Kuperman, V. & Bertram, R., “Moving spaces: Spelling alternation in English noun-noun compounds”. Language and Cognitive Processes, 28, 939–966, 2013.