Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 2 (2021) (pp77-85)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i2.206

Detection of Adjective Compound Word in Malay Language using Enhanced Syntactic Rules

Zamri Abu Bakar[1*], Normaly Kamal Ismail[2], Nurhilyana Anuar[3], Aminatul Solehah Idris[4]
[1,3,4] Centre of Foundation Studies, Universiti Teknologi MARA, Cawangan Selangor Kampus Dengkil, Malaysia
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam Selangor, Malaysia

*Corresponding author: * zamri@uitm.edu.my*
Received Date: 24 February 2021
Accepted Date: 10 March 2021
Revised Date: 21 March 2021
Published Date: 1 April 2021

HIGHLIGHTS

- Most of the languages consists of grammar rule's structure to construct a sentence.
- Enhanced syntactic rules approach was used to extract adjective compound word.
- Experimental precision result obtained is 70.3% which has increased 0.3%.
- Result was significant in terms of effectiveness dependency relationship.

ABSTRACT

Compound word is defined as combination two or more words and it will produce a new meaning. Generally, compound word is existed in many languages such as English, Mandarin, Arabic and others. Although, there are discussion of existing methods to detect compound word yet some limitations on detecting Malay compound word. Thus, this study is done to improve accuracy towards adjective compound words. Training data is used in this study was Malay story books. Digitization data of Malay story book is used in this study. Then, the pre-processing method involved tokenization, stemming, bi-gram and part-of-speech (POS) tagging has been applied to produce the candidate compound word. Applying the enhanced syntactic rules shown the precision result is 70.3% through this study. Thus, this study will contribute to the academic research in improvise the issues on searching and document summarization application.

Keywords: *Compound word, Syntactic rules, Malay language*

INTRODUCTION

Language is a central method for human to communicate to the community. Through language, it helps people to express their thought and feeling. In each country and region in this world, there have mother tongue, and formal language have been used in daily life. There are several languages have been used to communicate worldwide such English, Chinese, Arabic, Malay and others. Each language is unique, language help to portray the culture of the community or region. For example, most of Arabic people in middle east country such United Arab Emirates, Yemen, Messer, Jordan, Syria, Iraq, Iran and country nearby will use Arabic language as spoken and written. Thus, it portrays Muslim community. Furthermore, language also widely used in education, business, music, and others.

In most of the languages have systematic rule and structure to construct a sentence. Part of component in the language consist of grammar, word, compound word, sentence and others. What is compound word? Compound word is combination of two or more words. This combination of the words will produce new words and give new meaning. For example, combination of two noun words (noun+ noun) such as football. Studies in compound word is widely attested in many languages such as English, Mandarin, Arabic (Christianto, 2020; Gagné, Spalding & Schmidtke, 2019; Altakhaineh, 2016). In Malay language, the study has started around year 2011 by Rahman, Omar and Aziz on detecting head modifier of noun phrase in Malay sentences. It shows that limited studies have been explored in this language area. Furthermore, study conducted by Rahman, Omar and Aziz (2013) on extraction of compound word in Malay text shown that they focus on compound nouns in Malay noun phrases and the accuracy of compound noun results was evaluated by the linguist or the language expert to verify the correctness of the compound noun produced from the prototype system. In other words, there is still lack of studies discussing on adjective compound word in Malay language. Thus, this study will focus on improving accuracy of detection adjective compound word in Malay story book using syntactic rules. The important of this study is to contribute for several application such as text summarization, machine translation, Information Retrieval (IR), machine translation, and semantic analysis (Rahman et al., 2011; Rahman et al., 2013).

Next section will be discussed related work on compound word. Followed by methodology and final section will be result and conclusion.

RELATED WORK

A large number of researches actively have been discussed on compound word across different languages. Compound words generally existed across different languages were defined as it composed of at least two root-words (Shen, Li & Pollatsek, 2017). As cited in Cahyanti (2016) compound word has three forms: closed form, compounds written as single words (newspaper, goldfish, highway); the hyphenated form, compounds that are hyphenated (daughter-in-law, third-rate, court-martial); the open form, compounds written as separated words (red zone, high school, health care). This study has adapted lexical meaning and contextual meaning to identify compound words in Twilight novel written by Stephenie Meyer. Through this quantitative study, the results found 253 compound words which focus on written, world class and meaning perspective.

Many studies have discussed the detection of noun-noun compound words as compared to adjective-noun compounds. However, an empirical study has been attested in German language to find suitability of adjective-noun compound as naming unit or phrase (Schlechtweg, 2018). Another study discussed by Ang, Tan and He (2017) employed node-and-collocate approach to identify noun-noun and adjective-noun collocations in English language research article. Thus, result shown general adjective is the most common noun pre-modification type found in the articles.

Fundamental study conducted by Rahman et.al. (2011), identified Malay grammar mainly can be categorised into sound of the language and arrangement of word in a sentence. By applying concept thematic relation of compound words. Thus, the study has come out with two main compound noun phrase categories are i) head and noun modifier ii) head and non-noun modifier. Result from this study by using thematic relation, has come out with construction of four basic compound phrase such as Noun Phrase + Noun Phrase (NP+NP), Noun Phrase + Verb Phrase (NP+VP), Noun Phrase + Adjective Phrase (NP+AP) and Noun Phrase + Preposition Phrase (NP+PP) in field of Malay language.

Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 2 (2021) (pp77-85)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i2.206

METHODOLOGY

This study applied five phases as a method to process the detection of adjective compound word, Figure 1 is shown that the flow of the method used in this study. The phases include; (i) digitize five Malay story books as a corpus, (ii) pre-processing tasks consist of two tasks which are normalization and tokenization (iii) the candidate generation consists of subject and predicate phrase, POS tagging, word bi-gram and candidate collocations frequency equal or greater than two (iv) automatic compound word detection (syntactic rules) (v) the evaluation metric which is precision and recall is used to evaluate the efficiency of the enhanced method used.

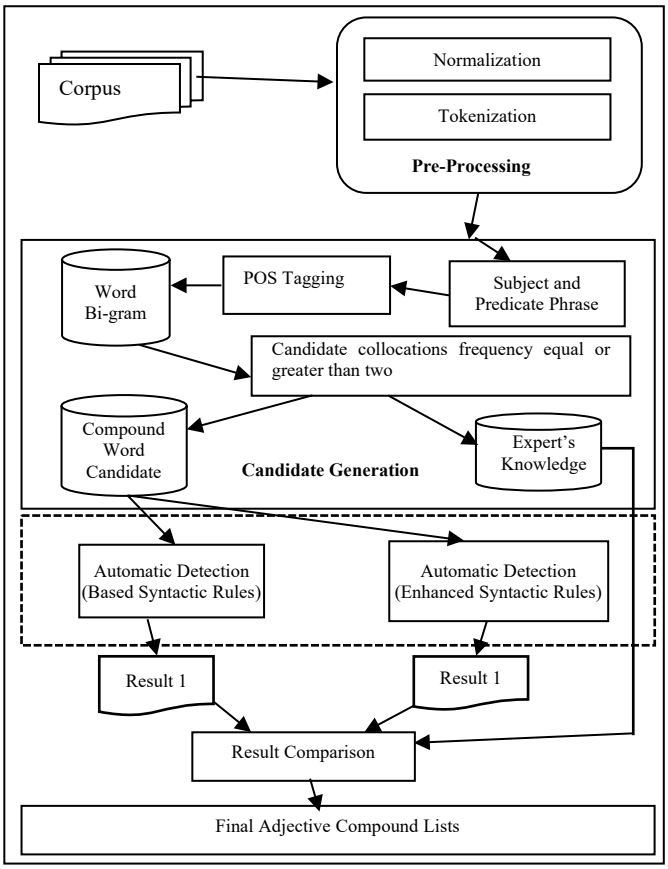

Figure 1: Detection of Compound Adjective Framework

Corpus

Corpora have been extensively employed in several Natural Language Processing (NLP) tasks as the basis for automatically learning models for language analysis and generation. In this step, this study collected the data from story book which is written in Malay language. The size of the corpus is 30 pages and 3,116 tokens.

Pre-Processing

In this phase, all the digitization of story book is pre-processed by removing all page's tags, header and footer tags, remove all noisy data and breaking the content down to a sequence of individual tokens. After that, all-uppercase, capitalized and mixed case words were lowercased. Punctuations, special symbols and numbers are removed. All this process is known as normalization of the data. After the normalization

process is done, we proceed to the next step which is tokenization. We develop the system using the Microsoft Office Studio to tokenize all the words in the digitize data. Table 1 shows the statistic of our Malay story book corpus.

Table 1: Statistics of the Malay corpus

Number of sentences	450
Number of tokens	3,116

Candidate Generation

In this phase, we have tagged all the word related with its POS tags in the text corpus that match with all Malay lexicon words in the database. After this POS tagging process is done, we construct the bi-gram word to all the word from the corpus. Table 2 shown the statistic of word bi-gram. This phase gives all possible Noun-Adjective (N-A) collocations that occur in a corpus. From the tagged corpus, if two consecutive words tagged as Noun and Adjective respectively is extracted as a candidate N-A collocation. These compound Adjective candidates are then passed to the next phase for automatic compound words extraction method. Compound words candidates which occur with frequency equal to one are not selected because more compound word will produce with irrelevant collocation between two words. Study conducted by Muneer et. al. (2016) stated compound word candidates with frequency in the corpus are greater than or equal to three are selected in their study. For this study, we have enhanced the syntactic rules which suitable with the dataset that we used from the story book where only candidate compound adjective collocations whose frequency in the corpus are greater than or equal to two are chosen.

Table 2: Statistics of the number of bi-grams

Number of bi-grams	2,046

Automatic Extraction

Once we have extracted the candidate N-A compounds in the compound word candidate generation phase, we have ranked the entire compound adjective according to higher frequency word bi-grams to lower word bi-gram from a corpus. In our task, linguistic approaches are used to get a valid compound. There are some experiments run to make sure the result produce with valid compound word and it verified by the expert.

Syntactic Rules

When we extract the compound words from the compound word candidate generation phase, this study proposes enhanced syntactic rules to detect the compound word adjective candidate from the Malay corpus. Thus, it means that, this study proposes linguistic knowledge approach to extract and categorize the compound adjective from the Malay corpus. In order to extract compound nouns in standard Malay sentences, we must understand the language grammar itself first. Basically, the grammar for Malay language describe that the sentence must have a subject, verb and predicate.

In this research, the first step to run the experiment is to separate all the sentence in the story into all the single sentence.

Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 2 (2021) (pp77-85)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i2.206

Story
Di pinggir sebuah hutan tinggal sepasang katak. Katak jantan sangat baik hati. **Katak betina** pula sombong dan tidak suka berkawan dengan **katak lain**.

List of the sentences
1. Di pinggir sebuah hutan tinggal **sepasang katak**.
2. **Katak jantan** sangat **baik hati**.
3. **Katak betina** pula sombong dan tidak suka berkawan dengan **katak lain**.

The next step is, we separate the sentence into subject and predicate. Below is a output after the process is done.

Output
1. **Katak betina** (Subject)
2. pula sombong dan tidak suka berkawan dengan **katak lain**. (Predicate)

The next 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.

Process
1. **Katak betina** (Subject)
2. ~~**pula**~~ sombong ~~**dan tidak**~~ suka berkawan ~~**dengan**~~ **katak lain**. (Predicate)

Output
1. **Katak betina** (Subject)
2. sombong || suka berkawan || **katak lain**. (Predicate)

The following are some examples of syntactic rules constructed based on the structure of the sentences.

Sentence
Katak betina pula sombong dan tidak suka berkawan dengan **katak lain**. (Female frogs are arrogant and do not like to be friends with other frogs.)

The POS tag is done for each word in this sentence below:

"*Katak[Noun]Betina[Adjective] pula[Modifier] sombong[Adjective]and[Conjunction] tidak[Modifier] berkawan[Verb] dengan[Conjunction] katak[Noun] lain[Adjective].*"

Evaluation

The Performance and Accuracy Measurement are described below;

$$\text{Precision} = \frac{X}{(X+Z)} \qquad (1)$$

$$\text{Recall} = \frac{X}{(X+Y)} \qquad (2)$$

where:

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

Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 2 (2021) (pp77-85)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i2.206

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:

$$\text{F1-Score} = \frac{2\ X\ Precision\ X\ Recall}{(Precision + Recall)}$$

RESULT AND DISCUSSION

Table 3 showed the result number of adjective compound words found by the linguistic expert and prototype system has been developed. This table showed, number of compound words found from five different story books. Meanwhile, this study is also shown the result in form of the graph. Table 3 is referred to the figure 2.

Table 3: Result of adjective compound word by story book

	Syntactic Rules (N+A)	Syntactic Rules (A+N)	N+A (Expert)	A+N (Expert)
Book 1	3	2	4	3
Book 2	4	5	5	7
Book 3	7	4	11	5
Book 4	5	1	2	1
Book 5	4	2	6	3
Overall	23	14	28	19

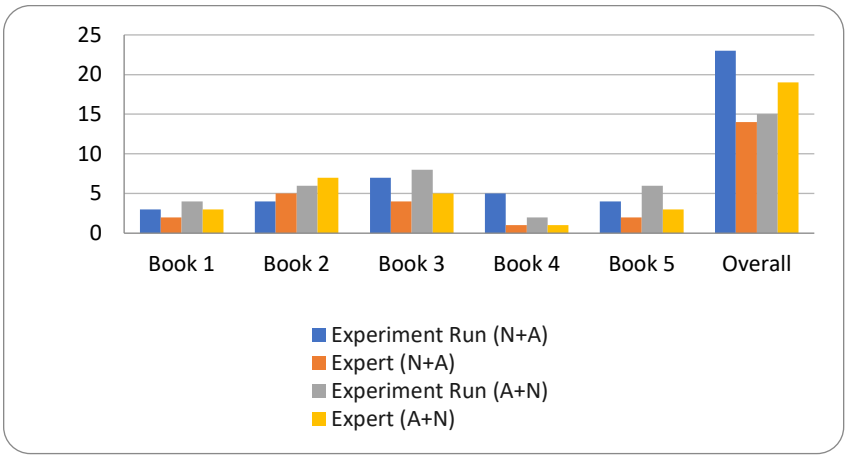

Figure 2: Result for Compound Adjective by Book

Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 2 (2021) (pp77-85)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i2.206

Table 4 showed the result number of compound words found by the linguistic expert and prototype system has been developed. This table shows, number of compound words found by each 100 sentences from the overall story book. Total of sentences from all the books are 409 sentences. Meanwhile, this study is also shown the result in form of the graph. Table 4 is referred to the figure 3.

Table 4: Result of adjective compound word by sentence

Sentence	Syntactic Rules (N+A)	Syntactic Rules (A+N)	N+A (Expert)	A+N (Expert)
100	5	1	9	2
200	9	2	11	4
300	2	1	4	1
400	1	0	1	0

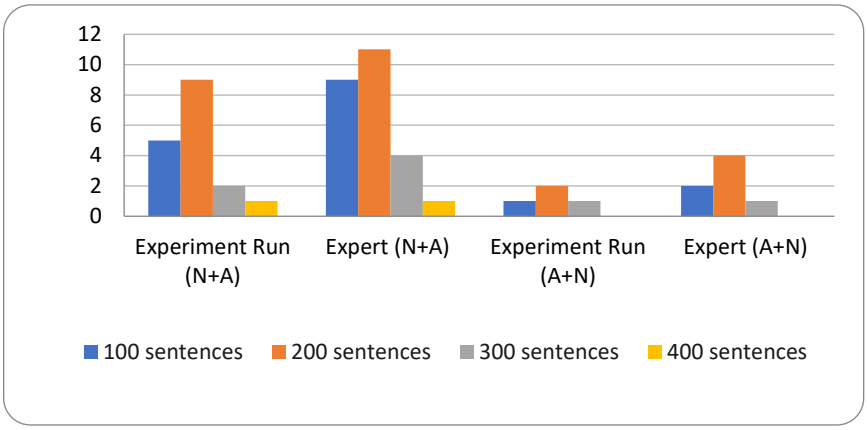

Figure 3: Result for Compound Adjective by Sentence

Table 5: Recall, Precision and F1-Score for Each Relation

Word Combination	Base-line			Enhanced Syntactic Rules		
	Precision	Recall	F-Score	Precision	Recall	F-Score
N+A	90.9%	10.2%	18.1%	91.2%	10.4%	18.5%
A+N	0	0	0	70.3%	5.7%	15.2%

The comparison of the results is shown in Table 5 showed the comparison of the result between based-line result and Enhanced Syntactic Rules. The result is measured according to the standard measurement which is precision, recall and F-Score in getting accuracy of the result. Finally, the recall, precision and F-Measure value by using Enhanced Syntactic Rules is showed that precision increased 0.3 percent, recall increased by 0.2 and F-Score increased from 18.1% into 18.5%. The increment of the percentage for the enhanced method is significant for this study even it is slightly lower. Thus, this study has also assisted in increasing the percentage values of improvement the result. Table 6 below is a few examples of adjective compound word for this study.

Table 6: Example of Adjective Compound Word

Compound Word Output	**Validation**
katak betina *(female frog)*	Yes
air besar *(big water)*	Yes

katak jantan *(male frog)*	Yes
pula sombong *(is arrogant)*	No
katak lain *(other frog)*	Yes
suka berkawan *(love to be friends)*	No

CONCLUSION

This study has discussed how the enhanced Syntactic Rule in a Malay language 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.

ACKNOWLEDGMENTS

This work is financially support by Universiti Teknologi MARA Cawangan Selangor Kampus Dengkil. The authors also want to express a special thanks to Administrative of Universiti Teknologi MARA for all the support.

CONFLICT OF INTERESTS DECLARATION

The authors declare no conflict of interests regarding the publication of this article.

REFERENCES

Altakhaineh, A. (2016). Identifying Arabic compounds other than the Synthetic Genitive Construction, Acta Linguistica Hungarica Acta Linguistica Hungarica, 63(3), 277-298. Retrieved May 25, 2020, from https://akjournals.com/view/journals/064/63/3/article-p277.xml

Ang, L. H., Tan, K. H., & He, M. (2017). A corpus-based collocational analysis of noun premodification types in academic writing. 3L: Language, Linguistics, Literature®, 23(1).

Cahyanti, R. D. (2016). Compound words used in Stephenie Meyer's Twilight. Journal on English as a Foreign Language (JEFL), 6(1), 59-70.

Christianto, D. (2020). Compound Words In English. LLT Journal: A Journal on Language and Language Teaching, 23(1), 27-36.

Gagné, C.L., Spalding, T.L. & Schmidtke, D. (2019) LADEC: The Large Database of English Compounds. Behav Res 51, 2152–2179 (2019). https://doi.org/10.3758/s13428-019-01282-6

Muneer, A.S.H., Omar, N., Ba-Alwi, F.M. & Albared, M. (2016). Automatic Extraction of Malay Compound Nouns Using a Hybrid of Statistical and Machine Learning Methods. International Journal of Electrical and Computer Engineering (IJECE). 6. 925. 10.11591/ijece.v6i3.9663.

Rahman S. A., Omar, N. and Aziz, M.J.A. (2011). A fundamental study on detecting head modifier noun phrases in Malay sentence," *2011 International Conference on Semantic Technology and Information Retrieval*, Putrajaya, 2011, pp. 255-259, doi: 10.1109/STAIR.2011.5995798.

Rahman S. A., Omar, N. and Aziz, M.J.A. (2011). Transformation of Malay head modifier noun phrase into a thematic relation structure. *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, 2011, pp. 1-5, doi: 10.1109/ICEEI.2011.6021798.

Rahman S. A., Omar, N. and Aziz, M.J.A. (2013). Extraction of compound nouns in Malay noun phrases using a noun phrase frame structure. *Asia-Pacific Journal of Information Technology and Multimedia*, 3.

Schlechtweg, M. (2018). The naming potential of compounds and phrases: An empirical study on German adjective-noun constructions. *Word Structure*. 11. 359-384. 10.3366/word.2018.0133.

Shen, W., Li, X., & Pollatsek, A. (2017). The processing of Chinese compound words with ambiguous morphemes in sentence context. *The Quarterly Journal of Experimental Psychology*, 1-10.