2012 International Conference on Information and Knowledge Management (ICIKM 2012)
IPCSIT vol.45 (2012) © (2012) IACSIT Press, Singapore

Parser with Sentence Correction for Malay Language (BM)

Yusnita binti Muhamad Noor and Zulikha binti Jamaludin*

School of Computing College of Arts and SciencesUniversiti Utara Malaysia
06010 Sintok, Kedah, Malaysia

Abstract. This paper discusses the design methods for Bahasa Melayu (BM) parser with sentence correction that begin with tokenizing, checking the words, tagging POS (part of speech), validating and matching of context-free grammar (CFG), and proposing the correction and giving output. To date, there are no developed systems that have been designed in processing BM language for this purpose. Therefore, it has led to the development of a prototype for this purpose in accordance with the methods introduced. Results from the prototype development have shown a good output in which it could identify a wrong sentence and suggest the correct one based on the rules provided.

Keywords: BM sentence parser, BM sentence correction

1. Introduction

The Malay language is a context free grammar where there is a subject and a predicate in a sentence [1]. It requires a set of grammar rules also known as context-free grammar (CFG) in English or phrase structure rules (RSF) in BM [2]. Every sentence used in a language is constructed according to the CFG, especially in BM. For this reason, there are lots of research have been conducted in language studies in producing a good sentence structure especially in BM. Sentence parser is one of the tools of technology that can be used in validating a sentence to produce a good sentence structure. The program is also known as a syntactic parser by some researchers. It parses the sentence according to the CFG provided. Its function is to validate the construction of words used in a sentence. If a sentence is structured according to the rules of CFG, the parser will classify the sentence as true.

There are some studies conducted by BM researchers on sentence parser as cited in [3], [4] and [5]. The studies could validate a sentence according to the CFG rules. To date, parser with sentence correction has not yet been established for any language, especially BM. Thus, this study is to take up this challenge in producing an algorithm in the development of BM parser with sentence correction. Methods involved in the design development process will be introduced.

The structure of the paper is as follows: Section 2 briefly introduces related studies in BM sentence or syntax parser. Section 3 will discuss about the proposed design method. Prototype development is shown in section 4 and finally, Section 5 concludes the paper.

2. BM Sentence Parser

Reference [3], [4] and [5] had conducted studies on the syntax parser for BM sentence. The studies showed the program could parse sentences following rules provided in the system; the output produced informed the user whether the input sentence are categorized correctly or not based on the theory of transformational-generative grammar.

BM parser by Rosmah in [3] focused on BM compound sentences only. It acts as a superset to a translation aid system carried out by *University Kebangsaan Malaysia* (UKM). CFG rules provided by

[*] Yusnita binti Muhamad Noor. Tel.: + (0192816740); fax: (-). *E-mail address*: s92715@student.uum.edu.my
Zulikha binti Jamaludin. Tel.: + (04-9284061); fax: (04-9284054/9284753). *E-mail address*: zulie@uum.edu.my

Dewan Bahasa dan Pustaka (DBP) were used in parsing the sentence. The first process performed by the system is analyzing the morphology by dividing the sentence into words and word classes. Each word class was tested by syntax parser whether the sentence meets the CFG rules to produce the output. Reading, checking, and parsing the sentence was done using top-down parsing method. If the syntax matches the CFG, then a parse tree is produced. Otherwise, a message error is displayed. The processes involved are shown in the example below:

Examples of input: *Cakera liut itu murah*

Word	**Word class**
Cakera liut	Noun (N)
Itu	Determinant (D)
Murah	Adjective (A)

Rules: Subject=NP
NP= N (noun)+A (Adjective)+Determinant
Predicate=AP (adjective phrase)/A /N
Retrieved input: S=N+D+A
Output: Subject: NP (*Cakera liut, itu*)
Predicate: A (*murah*)

Suzaimah as in [4] also conducted research in this area which focused on both single and compound sentence. The analysis was done by matching input with CFG in parallel using Parlog programming and applying top-down parsing method. The first process performed by the system was to make a match for each word with the lexicon. Each input string was connected to the main classification and sub-classification prescribed in the lexicon. If the input sentence did not meet the CFG, then the error message was produced. Input must be in the form of Parlog code which required sentences to be separated by commas and parentheses.

For example, input [*ali, makan*] will produce the following output:
P=ayat(frasa_nama(ung_nama(ali))),frasa_kerja(ung_kerja(kata_kerja(makan))) succeeded.

Ahmad Izuddin as in [5] also conducted research in BM sentence parser. The prototype produced analyzed syntax and semantic for sentence entered by user. For correct sentence structure according to CFG will produce a tree diagram. Otherwise, it displayed an error message. The processes involved in analyzing the sentence are similar to Rosmah' parser as in [3]. Experimental results showed that the prototype was able to correctly parse the provided sentences by approximately 81%.

2.1 Comparison of BM sentence parser

From the previous studies as described above, parser in [3] and [5] can be seen as to have more in-depth studies compared to Suzaimah's parser as in [4] because the study produced output in the form of a syntax tree and receiving input was in the form of sentence. Compared with Suzaimah' parser, the limiting factor was the input which was only in Parlog code. Besides, the resulting output (Parlog clause) is hard to understand by some users. The similarities and differences from the studies are described in Table 1.

Table 1: Comparison of BM Sentence Parser

Similarities	Differences		
Rosmah's parser [3], Suzaimah's parser [4], Ahmad Izuddin's parser [5]	Rosmah's parser	Suzaimah's parser	Ahmad Izuddin's parser
1. The applications will match each word with the lexicon to validate the order of words or word classes according to the rules 2. If it fits the rules, either output (syntax tree or Parlog) clause or error messages will be produced 3. The applications also use top-down approach which is a recursive descent parser.	1. Input in sentence form 2. Output produced in parse tree 3. Weaknesses: • Cannot distinguish between the subject noun phrase because the existence of description in the CFG • The subject must end with a determinant • The sentence entered by the user should not contain punctuation mark such as commas, exclamation points or other punctuation • The system does not cater the fundamental problem in parsing, such as the POS ambiguity [8].	1. Input in Parlog code 2. Output produced in Parlog clause 3. Weaknesses: • Input sentence should be in Parlog format and for those who do not understand the format will find it difficult to use the system • The study also did not use the rules of CFG approved by DBP or any BM experts, but it was developed according to the suitability of the sentences itself.	1. Input in sentence 2. Output in parse tree/tree diagram 3. Weakness: • The prototype could not identify any single error of grammar's component.

3. Design Methods for BM Parser with Sentence Correction

Study as in [5] did not explain the type of method used. However, methods used in receiving input and producing output were considered as having the same methods used by Rosmah as in [3]. Therefore, in the development of a prototype for this study, the methods used in [3] and [4] will be followed and extended. Both studies analyzed the sentence by 1) dividing the sentence into words, 2) assign word classes, 3) syntax matching and 4) output. In this study, different output structure will be produced. The order of CFG syntax and syntax tree visualization will be produced when the sentence conform to the correct structure. Otherwise, an algorithm to give suggested corrections will play a role.

To date, system that can propose a correction in terms of sentence structure has not yet been created for any language, especially BM. Therefore, the proposed method in correcting a sentence as done in this study will contribute new ideas in language-based studies, especially BM. Sentences which cannot be analyzed by the system due to the incorrect use of the language and the use of severely wrong sentence structure according to BM CFG will not issue the proposed correction, but an error message which is displayed so that the user will enter the sentence again.

As a beginning, this study will be limited only to a single sentence not exceeding 14 words per sentence as was done in [3] and [4] who focused on particular types of sentences.

The design methodology involved in BM parser with sentence correction are 1) token sentence into words, 2) check the number of words, 3) POS tagging, 4) Check spelling or check the conjunctions, 5) validation and matching of CFG and 6) suggestion or visualization. Each process will be expanded into sub-processes, as illustrated in Fig. 1:

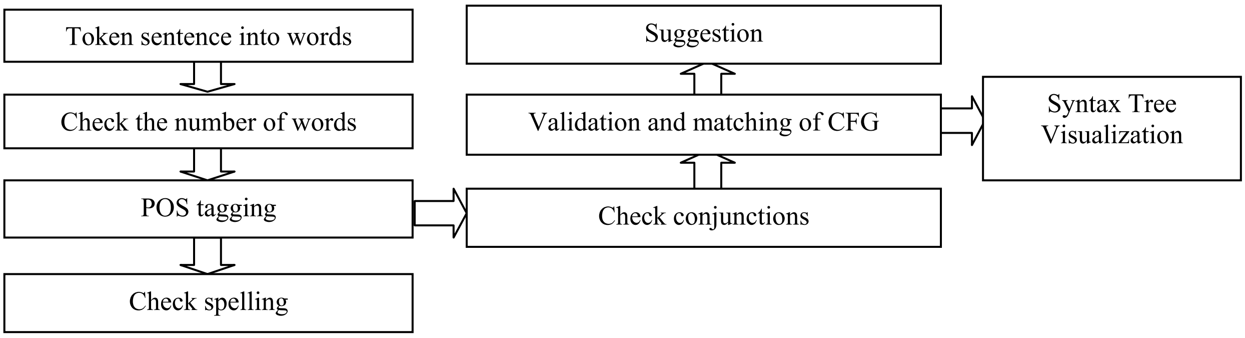

Token sentence into words
Check the number of words
POS tagging
Check spelling
Suggestion
Validation and matching of CFG
Check conjunctions
Syntax Tree Visualization

Fig. 1: Design methods for BM parser with sentence correction

Processes involved in Fig.1 are described below:

1. Token sentence into words
   After input is received, the sentence will be token into words.
2. Sentence conditions are reviewed to ensure that the input sentence is more than one word.
3. Each word will be matched with the appropriate POS as provided in the repository.
4. If there are words that cannot be matched, then the spell checker will play a role. Normally, there are certain words that are not listed in the repository like special noun words, date, address and numbers. Otherwise, sentence condition will be checked. It is to determine that the sentence is not a compound sentence by checking the conjunctions. If there is a conjunction in the words list, the system will produce an error message because only single sentence will be processed.
5. Determination of the validity in the input sentence is determined by matching the structure of input sentence or the order of word classes with CFG. CFG produced as in [6] is used as a reference. Matching success will continue to display the output. Otherwise it would require the proposed method of correction to be carried out.
6. Suggestion
   At this stage, similar CFG for input sentence will be searched according to the order of word classes listed for input and CFG. Only one similar CFG will be taken. Replacement will be done by changing the position of the words (input sentence) according to the word class order in CFG that have been retrieved. Hence, the proposed sentence will be displayed to the user. However, for

sentences that are too difficult to change, it will only give an error message (refer Fig. 2 for methods used in sentence correction).

7. Output
   For the correct sentence determined by the parser, it will display output in CFG (order of CFG) and syntax tree visualization. Method in display the output in syntax tree visualization will be discussed in more detail in future work.

3.1 Process of analyzing and recommending sentence correction

After checking the sentences by matching the order of word class (input) with the arrangement of CFG, sentences that do not match with CFG will go through the process of replacing. Similar CFG to the arrangement of word class for input sentence is achieved and replacement process will be performed by changing the position of words based on the CFG. A new sentence recommendation is displayed to the user for the purpose of input and conversion. Fig. 2 shows the process involved in making recommendations to the user. Examples of BM sentence that do not have the correct structure "*saya nasi makan*" is used.

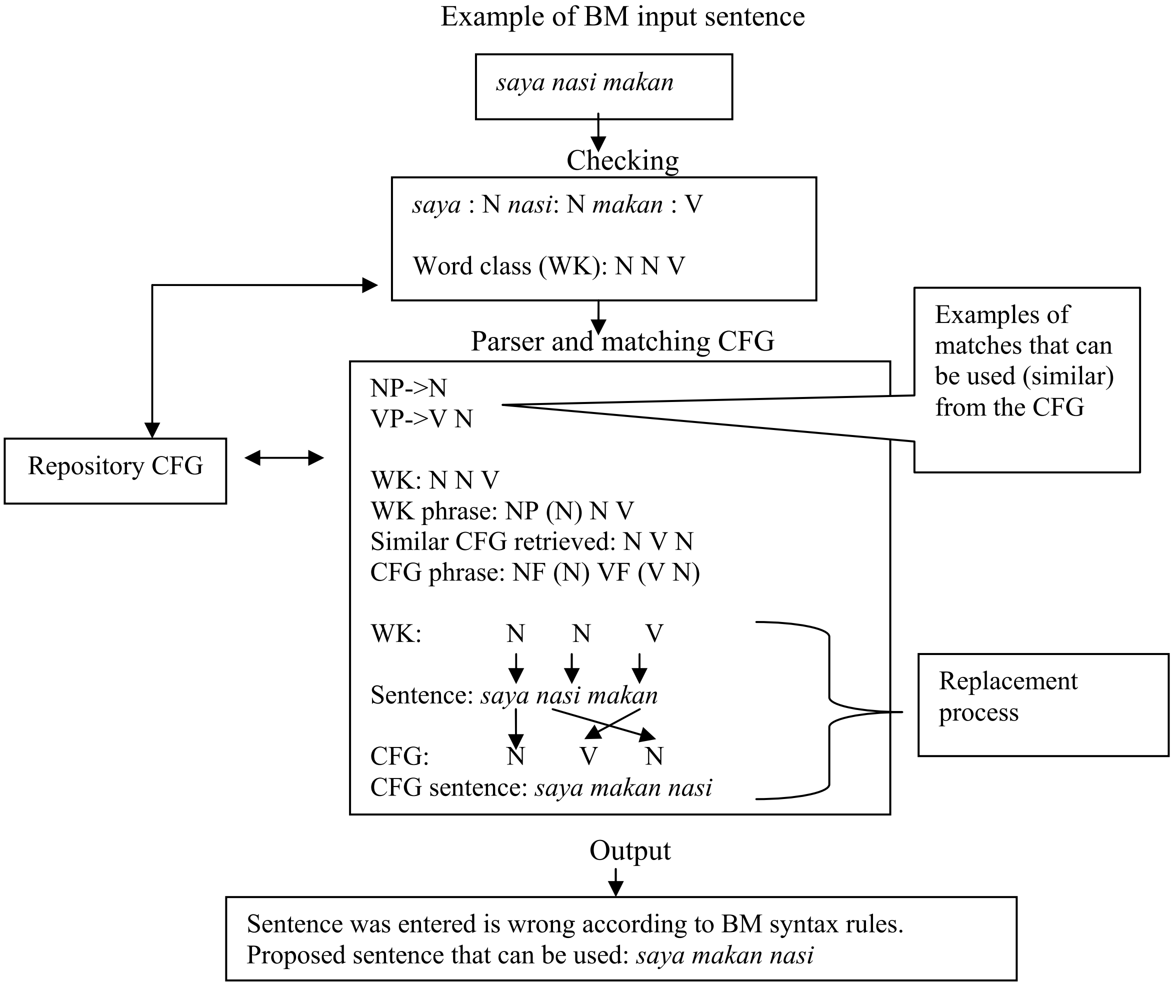
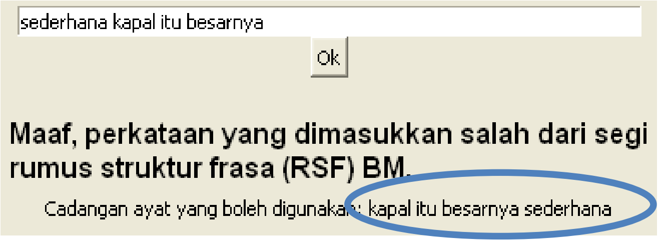

Example of BM input sentence
saya nasi makan
Checking
saya : N nasi: N makan : V
Word class (WK): N N V
Parser and matching CFG
Examples of matches that can be used (similar) from the CFG
Repository CFG
NP->N
VP->V N
WK: N N V
WK phrase: NP (N) N V
Similar CFG retrieved: N V N
CFG phrase: NF (N) VF (V N)
WK: N N V
Sentence: saya nasi makan
CFG: N V N
CFG sentence: saya makan nasi
Replacement process
Output
Sentence was entered is wrong according to BM syntax rules.
Proposed sentence that can be used: saya makan nasi

Fig. 2: Process of analyzing and recommending correct sentences

4. Result From the Prototype Development

The prototype for this study is still under development in which it also includes other components like BM word attributes and syntax tree visualization. However, the proposed parser with sentence correction has been able to be analyzed to determine the output results. The output measured in the development process was found to give a good output and meets the correct sentence structure. For example, Fig. 3a and Fig. 3b below are some examples of the results obtained for the incorrect sentence for BM.

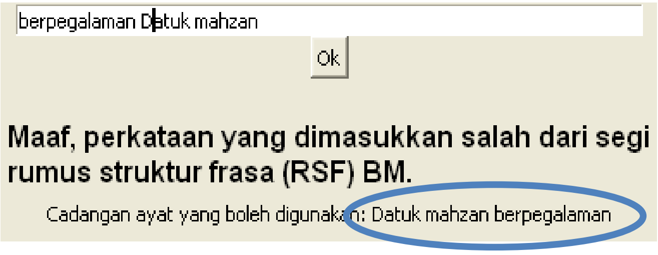


sederhana kapal itu besarnya

Maaf, perkataan yang dimasukkan salah dari segi rumus struktur frasa (RSF) BM.

Cadangan ayat yang boleh digunakan: kapal itu besarnya sederhana

Fig. 3a: Proposed correction of incorrect sentence



berpegalaman Datuk mahzan

Maaf, perkataan yang dimasukkan salah dari segi rumus struktur frasa (RSF) BM.

Cadangan ayat yang boleh digunakan: Datuk mahzan berpegalaman

Fig. 3b: Proposed correction of incorrect sentence

In Fig. 3a, the incorrect BM sentence structure entered is "*sederhana kapal itu besarnya*" or "medium the ship size" will give suggestion to " *kapal itu besarnya sederhana*" mean that " the ship have a medium size". In Fig. 3b, the incorrect BM sentence structure entered is "*berpengalaman Datuk mahzan*" or "experienced Datuk mahzan" will give suggestion to "Datuk mahzan berpengalaman" mean that " Datuk mahzan experienced".

5. Conclusion

Researches in the language field have grown in popularity among researchers from various countries. Therefore, the proposed design method in this study serves as a contribution to undertake studies related to language processing for any language. Parser introduced can be used in others language-based studies as a sub-tool needed such as in semantic processing, machine translation and etc. It also will be useful in sentence parse tree visualization as to check a grammatical sentence to be visualized. Even though research in BM processing is still at a moderate level compared to other languages [7], there would be a chance for BM to become one of the language focus by researchers similar to English if studies continue to be carried out.

Therefore, it can be concluded that this study has highlighted the design methodology for BM parser with sentence correction. Related studies with comparison have also been described. Process of analyzing and making recommendations to the incorrect sentence is also described. Illustrations showing in detail the processes involved in giving the suggestion are shown in Fig. 2. Example results from a prototype display are also included for reference. This study in turn will be expanded to create models and algorithm for BM syntax tree visualization with word attributes display in further studies.

6. Reference

[1] M. J. Ab Aziz et al. "Pola Grammar Technique to Identify Subject and Predicate in Malaysian Language," in *Proc. The Second International Joint Conference on Natural Language Processing*, 11-13 October 2005, pp. 185-190.

[2] M. J. Ab. Aziz. (2007). Pengkomputeran Linguistik Bahasa Malaysia [Online]. Retrieved Dec 28, 2010, Available: http://www.ftsm.ukm.my/programming/prosiding-atur07/08-Juzaiddin.pdf

[3] R. Abdul Latif, "Penyemak Sintaksis Ayat Bahasa Malaysia," M.S. thesis, Universiti Kebangsaan Malaysia, Bangi, Selangor, 1995.

[4] S. Ramli, "Reka bentuk dan implementasi suatu penghurai bahasa Melayu menggunakan sistem logik selari," M.S.thesis, Universiti Putra Malaysia, Selangor, 2002.

[5] I. Zainal Abidin et al. "Utilizing top-down parsing technique in the development of a Malay language sentence parser, " in *Proc. of the 2nd International Conference on Informatics*, 2007, pp. 128-134.

[6] N. S. Karim et al.. *Tatabahasa Dewan Edisi Ketiga*. Dewan Bahasa dan Pustaka: Kuala Lumpur, 2009.

[7] Z. Mohd Don, "Processing natural Malay texts: A data-driven approach." *TRAMES: A Journal of the Humanities & Social Sciences*, Vol. 14(1), p90. 2010.

[8] M. Z. Ahmad. Nazri et al. (2008). An exploratory study of the Malay text processing tools in ontology learning [Online]. Retrieved June 02, 2012, Available: http://webs.cs.utm.my/images/stories/content/publication/pars2008/session5/graphic/2.zakree.pdf