The Effectiveness of Bottom Up Technique with Probabilistic Approach for A Malay Parser

Muhammad Azhar Fairuzz Hiloh
abukhairul@gmail.com
Department of Accountant General
Ministry of Finance, Putrajaya

Mohd Juzaiddin Ab Aziz
din@ukm.edu.my
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia

Lailatul Qadri Zakaria
lailatul.qadri@ukm.edu.my
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia

ABSTRACT

Parsing is a process of analyzing the input string in a sentence to define the syntax structures according to rules of grammar. This task is performed by a parser which will produce a parse tree as output. However, a problem occurs when the parsing process produces two or more parse trees in which the parser unable to represent a precise parse tree. This limitation is caused by ambiguity in the structure of sentences. Ambiguity is occurred when a word is classified more than one category of syntax and its usage will affect the semantics of the sentence. Thus, the parser needs to have an approach to solve the ambiguity problem and is able to process the most appropriate parse tree to present a sentence. Like other languages in the world, Malay language, a national language for Malaysian, is not exempted from ambiguity problem. However, due to its grammar being context-free grammar, the probabilistic context-free grammar approach can be used to support the parser in determining a more accurate parse tree. This study focuses on the development of statistical parser using a bottom-up technique for Malay language. The training data, in the form of simple Malay language sentences, are collected from various sources. Based on this training data, a statistical lexical corpus of Malay language which consists of vocabulary, grammar rules and their probability was developed. The bottom up parsing will be supported by implementing Cocke–Younger–Kasami (CYK) algorithm. The parser's performance is evaluated based on its effectiveness to overcome ambiguity by suggesting a more precise parse tree. In conclusion, the Malay Language Parser can be useful to help user identify the appropriate parse tree and solve ambiguity issues in Malay Language.

Keywords: parsing technique; probabilistic context-free grammar; CYK; ambiguity; Malay language

INTRODUCTION

Malay language is a formal language that is widely used in administrative, education and business in Malaysia. Malay language has attracted many researchers to perform Natural Language Processing (NLP) studies both in linguistic and computerization (Sabrina et al., 2011; Ahmad et al., 2007; Rozana et al., 2011; Yusmita & Zulaikha, 2011; Noor Hafhizah & Karim, 2012; Nik Safiah, 1975). NLP is a field of artificial intelligence that aims to get

computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech (Jurasky & Martin, 2000). NLP implementations can be divided into several components, which are phonology, morphology, syntax, semantic, discourse and pragmatic. Phonology is the study the way sounds are organized in a language. Morphology concerns on the word forms. Syntax is the study of how words are put together to form correct sentences. Semantic is about analyzing meaning, what word means, and how these meaning combine in sentence to form sentence meaning. Discourse concerns on how the immediately proceeding sentence affect the interpretation of the next sentence and finally, pragmatic describes a relationship of meaning to the goals and intentions of the speaker.

This paper concerns on syntax or also called as parsing. Parsing is one of the most important step in NLP which enable machine to identify not just a part of speech for each word in a sentence such as noun, verb and adjective, but also for modeling constituent, a group of word sharing a lexical or phrasal category such as noun phrase, verb phrase and adjective phrase. Parsing is a process of analyzing the input string in a sentence to define the syntax structures according to the rules of grammar. Several studies have been conducted to develop a parser for Malay language by using a context-free grammar (CFG). CFG is one of the most commonly used system for modeling constituent structure in natural language. It consists of a set of rules or productions, each of which expresses the way that symbols of the language can be grouped and ordered together, and a lexicon of words and symbols. Furthermore, Knowles and Zuraidah (2006) has featured Malay language syntactic as Subject-Verb-Object (SVO) which is a basic pattern in the parser. However, like other languages, ambiguity in Malay language is inevitable. For example, the sentence "*Dia selak helaian dokumen itu*" (He/She is flipping the document pieces) will produce two parse trees as shown in Fig. 1.
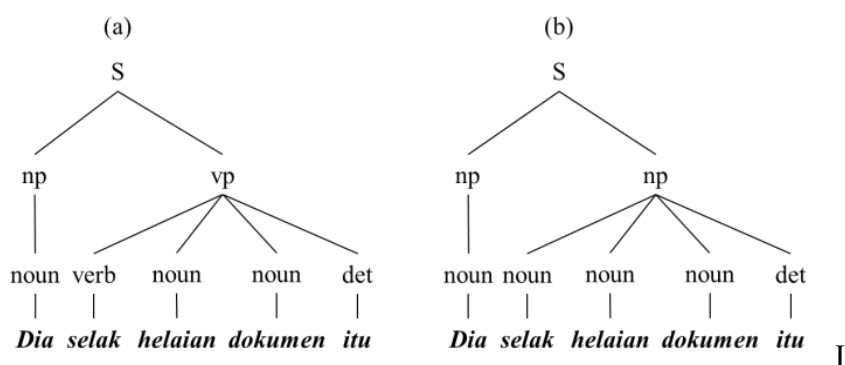

FIGURE 1. Two parse tree can be derived by a sentence "Dia selak helaian dokumen itu"

As explained in Charniak (1997), Dale (2000), Hassan et al. (2015), Nor Hafhizah (2011) and Jurafsky and Martin (2000), producing more than one parse tree is a common problem referred as an ambiguity, lexical ambiguity and semantic ambiguity, which leads to syntactic ambiguity. In this example, the word "*selak*" (flip) has more than one meaning, which is referred as semantic and lexical ambiguity. As a consequence, the word "selak" could be tagged with two different part-of-speech (POS) which are "*selak*" (flip) as an action tagged as a verb, and "*selak*" (latch) as a tool used to lock tagged as noun. Syntactically, two parse trees are produced to represent the sentence's structure, or called syntactic ambiguity. However, semantically parse tree (a) in Figure 1 is more precise representing the sentence. One of the common approaches to overcome ambiguity is by giving probabilistic value to the parse tree. According to Charniak (1993), the easiest mechanism to build a statistical parser is

by probabilistic context-free grammar (PCFG) where each context-free grammar rules is given a probabilistic value.

There are two types of parsing methods, top down parsing and bottom up parsing. Syntax analyzers follow the production rules defined by CFG and the production rules are implemented. When the parser starts constructing the parse tree from the start symbol and then tries to transform the start symbol to the input, it is called top-down parsing while the bottom-up parsing starts with the input symbols and tries to construct the parse tree up to the start symbol. To date, a small number of researches have been contributed to the development of Malay language parser in top-down parsing techniques while the bottom-up parsing technique is yet to be found. In addition, the development of Malay language parser incorporated with statistical approach is still in infancy. Thus, the aim of this study is to develop and evaluate the effectiveness of bottom-up parsing, supported with statistical approach, for Malay language parser. The bottom up parsing will be supported by implementing Cocke–Younger–Kasami (CYK) algorithm. CYK algorithm is a standard dynamic programming algorithm for parsing probabilistic context-free grammars (PCFGs).

RELATED WORK

Most of the studies involved in Malay language parsing were focusing on non-statistical parser with top-down technique. In Ahmed et al. (2007), a top-down parser was built to recognize whether a sentence consist of grammar or semantic error by classifying the Malay words into two categories of human and animal. The parser consisted of 3000 words gathered from Hawkins (1997) and grammar rules from Zulkifley et al. (2015), Nik Safiah (1995) and Nik Safiah et al. (2004). A difference parser's aims was developed by Rozana Kasbon et al. (2011), where the Malay language parser was functioned to translate short-form usage of Malay words into their correct spelling and suggest sentence correction if it contains grammar error. An enhanced Malay language parser was introduced by Yusnita and Zulaikha (2012) where the parse trees was visually represented and also included grammar correction. An early study on Malay language statistical parser was developed by Noor Hafhizah (2011) where the probabilistic approach was attached with top-down parsing technique. The system consisted of almost 40 thousand words based on the 2nd Edition of Oxford Dictionary. A set of training data with 147 grammar rules with their probabilistic value derived from 1000 simple Malay sentences was used in Noor Hafhizah (2011) to develop a statistical parser of Malay language. Noor Hafhizah gathered 90 words that have more than one part of speech tagging. Based on the review, most of the researches were using top-down approach while the bottom-up parsing technique was yet to be found. Therefore, to fill the gap, this research focuses on developing Malay Language Parser by implementing the bottom up with probabilistic approach.

METHODOLOGY

This study has been conducted in three main phases: (1) the development of Malay language corpus, (2) the development of statistical parser prototype, and (3) the evaluation of the prototype. Each phase is explained at the following sections.

THE MALAY LANGUAGE CORPUS

A statistical Malay language corpus has been constructed for the Malay language parser development. A dataset consists of 1700 simple Malay sentences has been collected for this study. The dataset is divided into two none-intersect subsets, following the approach introduced by Resnik and Lin (2013), with the ratio of 9:1 for the purpose of training and

testing data. Consequently, 1530 Malay sentences have been trained to construct a statistical Malay corpus. There are four main processes in constructing the statistical Malay corpus which are data collection in the form of simple Malay sentences, Part-of-Speech (POS) tagging for each word in the sentences, grammar rule assignment to each sentence, and finally probabilistic value calculation to each word with its POS as well as sentence with its grammar rule. There are a number of researchers working on Malay POS tagging such as Hassan Mohammad et al. (2011) and Nur Ashikin and Nazlia (2017). The Malay language sentences in the training data are categorized based on their constituent into four main grammar patterns which are *Frasa Nama* (Noun Phrase), *Frasa Kerja* (Verb Phrase), *Frasa Adjektif* (Adjective Phrase) and *Frasa Sendi* Nama (Preposition Phrase). The grammar patterns and their rules are constructed based on Nik Safiah et al. (2008) and Nik Safiah Karim (1975). The probabilistic value is calculated using (1) according to Collins (2003). Table 1 shows the descriptions for tokenization, part of speech tagging and constituent processes. The part of speech tagging and constituent identification process were done manually by linguistic experts to minimize the tagging error.

$$P(\beta \mid X) = \frac{Count(X \rightarrow \beta)}{Count(X)} \tag{1}$$

TABLE 1. Descriptions for main processes in the Malay Corpus development

Stages	Descriptions	Outputs
Sentence	Input	Sungai itu jauh dari rumah mereka.
Tokenization	Identify each word in the sentence	Sungai, itu, jauh, dari, rumah, mereka
Part of Speech Tagging	Identify word classes for each word.	Sungai/KN, itu/Pent, jauh/KA, dari/ KSN, rumah/KN, mereka/KN
Constituent / Phrases	Identify group of words that may behave as a single unit/phrase.	FN(sungai/KN +itu/Pent), KA (jauh/KA+ dari/KSN + rumah/KN + mereka/KN)

MALAY LANGUAGE SENTENCE PARSER'S ARCHITECTURE

As reported in Dale (2000) there are five phases in NLP which are tokenization, lexical analysis, syntactic analysis, semantic analysis and pragmatic analysis. Our study consists of three early phases and presented in the architecture as shown in Fig. 2. To evaluate the effectiveness of the proposed parsing technique and approach, a parser is developed following the prototyping technique described in Alavi & Umanath (1989) and Boar (1984). The system's input is a simple Malay language sentence. The output is a parse tree, or list of parse trees, that represents the sentence syntactic structure. If a sentence has more than a parse tree, the parser will propose the most precise parse tree.
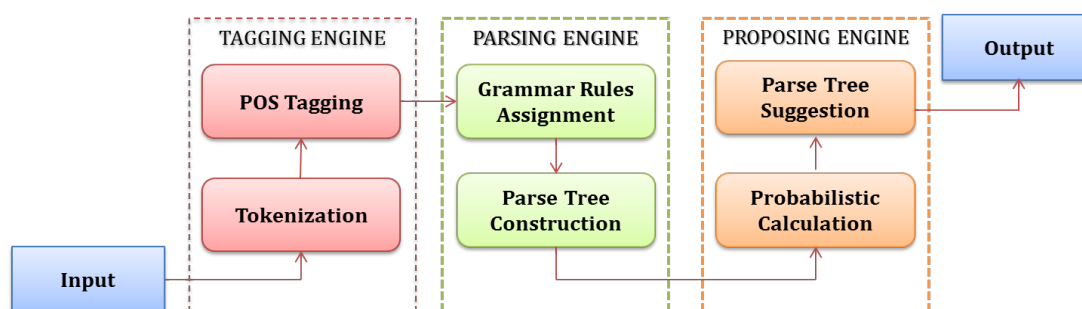

FIGURE 2. Malay Language Sentence Parser's Architecture

Mainly, there are three engines to support the architecture. The engines are explained at the following sections:

1) *Tagging engine:* In accordance to Dale (2000), an input sentence first needs to be separated and words are tagged with their POS individually. As explained in Palmer (2000), the tokenization process is to detach the sequent strings by determining their border. Thus, the word is matched and tagged with its POS by referring to Malay language corpus as shown in Fig. 3. In this example, each word in the given sentence "Saya selak pintu itu" was tagged with "saya/Kata Nama/KN", "selak/Kata Kerja Transatif/KKTR", "selak/Kata Nama/KN", pintu/Kata Nama/KN", itu/Kata Nama/KN". The tagging engine has tagged the word "selak" with more than one part of speech which are KKTr and KN.
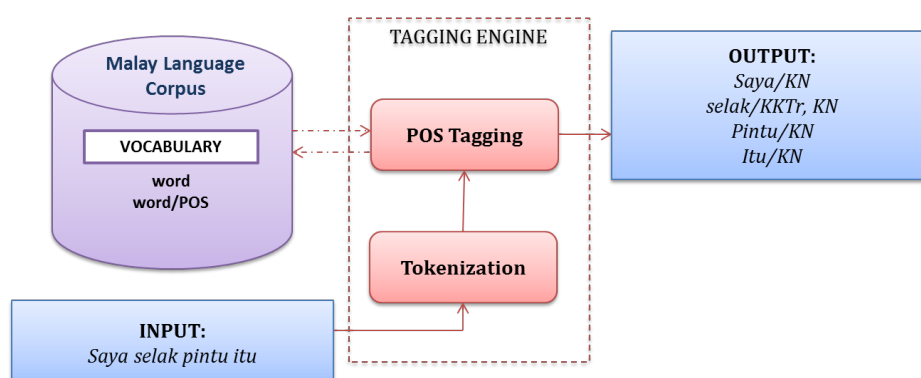

FIGURE 3. The tagging process

Parsing engine: The sentence syntactic structure is analyzed in two main processes at parsing engine which are assigning the grammar rules, and constructing the parse tree (Fig. 4). Both processes are performed by bottom-up parsing technique with Cocke-Younger-Kasami (CYK) algorithm. Bottom-up parsing is away from left-recursion problem and inefficient reparsing of subtrees which occur in top-down parsing as claimed in Jurafsky & Martin(2000) making this technique chosen in our architecture. Hopcroft et. al (2006) has described that CYK algorithm includes the following foundations: 1) applying dynamic programming known as table-filling and, 2) the grammar rules structure must be in CFG. Fig. 5 shows an example of the CFG grammar rules required to analyse the given sentence "Saya selak pintu itu".

A → S + P	FN → KN + KN + KN
S → FN	FN → KN + KN + Pent
S → FK	FK → KKTr + KN + Pent
P → FN	Saya → KN
FN → KN	Selak → KN
FN → Pent	Selak → KKTr
FN → KN + KN	Pintu → KN
FN → KN + Pent	Itu → Pent

FIGURE 4. Example of a set of CFG grammar rules for Bahasa Melayu

Based on Hopcroft et. al (2006) explanation, the general table-filling foundation involved constructing tabulation as shown in Fig. 5. The horizontal axis in Fig. 6 shows the input position $w = a_1 a_2 \ldots a_n$, where *w* is the input sentence that contains *n* words. The

sentence "saya selak pintu itu" contains four words, thus n = 4. The table-filling is done level by level, from *Level₁* to *Levelₙ*. Rows and columns of the table correspond to the start and end positions of a span. A cell in the table corresponds to the sub-string that starts at the row index and ends at the column index. It contain information about the type of constituent (or constituents) that span(s) the substring, pointers to its sub-constituents, and/or predictions about what constituents might follow the substring. Once this process is completed, the sentence is recognized by the grammar if the entire string is matched by the start symbol (A). Output for the parsing engine is the list of parse trees derived from the syntactic analysis. In this example, four parse trees are suggested by the system. The next procedure is to determine the most appropriate parse tree based on the grammar rules and probability calculation.
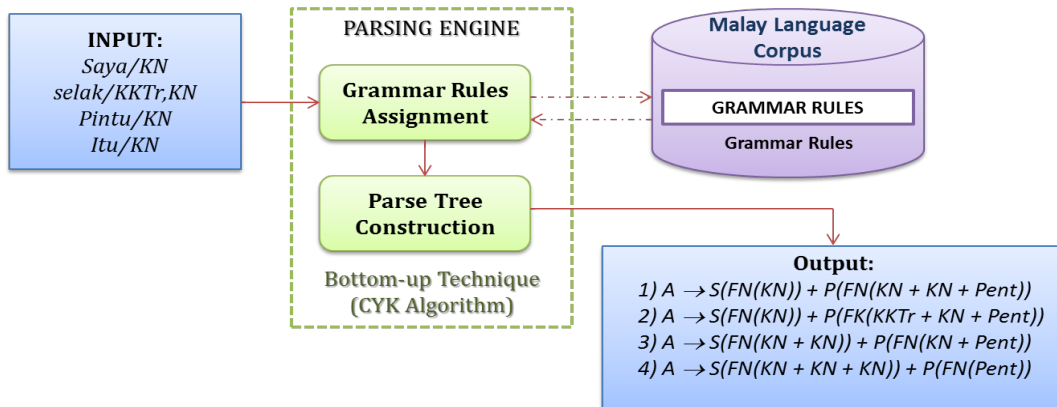

FIGURE 5. The parsing processes

Level 4	A→S(FN(KN))+P(FK(KKTr+KN +Pent)) A→ S(FN(KN))+P(FN(KN+KN +Pent)) A→ S(FN(KN+KN))+P(FN(KN +Pent)) A→ S(FN(KN+KN+KN))+P(FN(Pent))			
Level 3	(KN + KN + KN) FN → KN + KN + KN S → FN(KN + KN + KN)	(KN+KN+Pent,KKTr+KN+Pent) P → FK(KKTr + KN + Pent) P → FN(KN+ KN + Pent)		
Level 2	(KN + KN) FN → KN + KN S→ FN(KN + KN)	(KN+KN, KKTr+KN) FN = KN + KN FK = KKTr + KN	FN = KN + Pent	
Level 1	(KN) FN → KN S→ FN(KN) P→ FN(KN)	(KN, KKTr) FN→ KN P → FN (KN)	(KN) FN → KN P → FN (KN)	(Pent) FN → KN P → FN(KN)
	Saya	Selak	Pintu	Itu

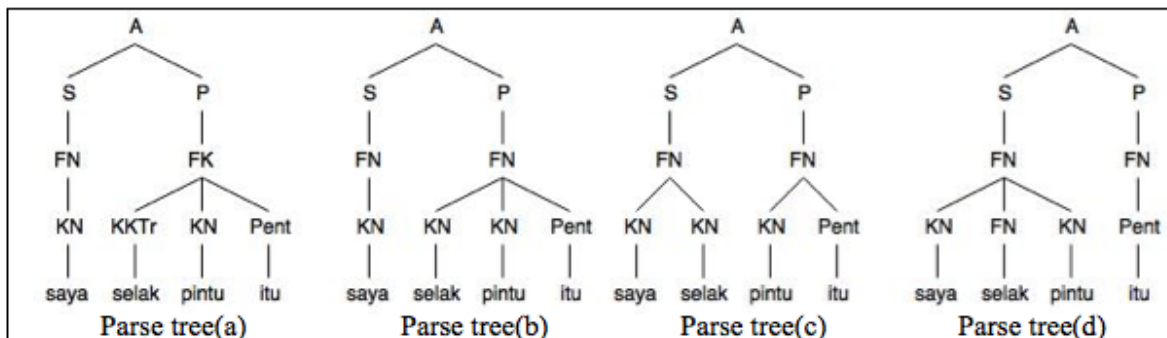FIGURE 6. CYK algorithm table filling for "Saya selak pintu itu"

FIGURE 7. Four parse trees were suggested from the sentence"saya selak pintu itu"

Proposing engine: Syntactic relationship is a method to describe the relation amongst words and phrases to form sentence in Malay language as reported in Hashim (1990). Furthermore, Chomsky (1980) describes that this relationship could be visually represented by a parse tree. In reference to Jurafsky and Martin (2000), the parsing process is referring to the analysis of the input sequence in a sentence for the purpose of determining its syntactic structure according to grammar rules. Likewise in other languages, the occurrence of ambiguity in Malay language is predictable. As shown in Fig. 6, four parse trees are derived from the input sentence. Therefore, the support of statistical approach has been adopted in our parser to provide its ability in proposing a more precise parse tree. The probabilistic value for each tree is calculated and the proposing engine will suggest the tree with the highest probabilistic value to represent the sentence, as shown in Figure 8.
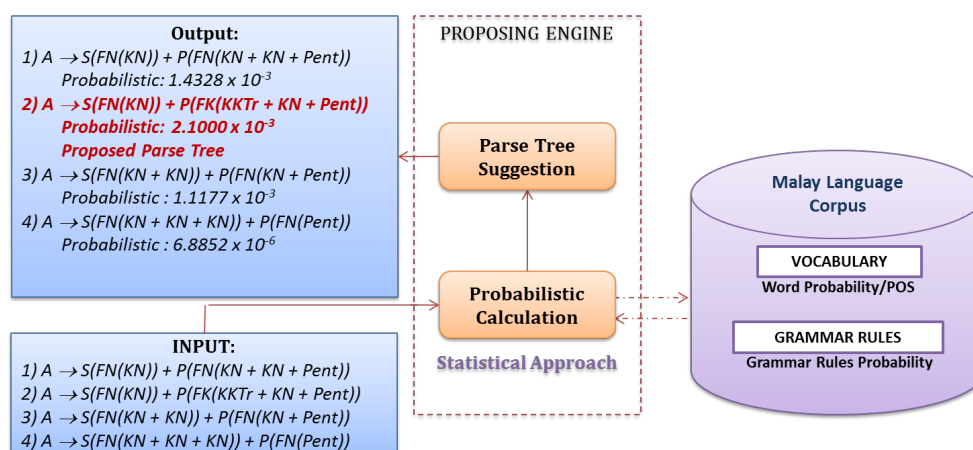

FIGURE 8. The proposing processes

EVALUATION RESULT AND DISCUSSION

In order to measure the parser's performance, a testing dataset, which contains 170 simple Malay language sentences, is obtained. Those sentences are randomly chosen from the overall 1700 sentences gathered in the dataset. The testing dataset, as principally mentioned in [9], is separately kept and never been visited until the parser development has completed. The parser performance is evaluated with the same measure in Ahmad et al. (2007) and Rozana et al. (2011) using the average as in (2) and weighted average as in (3):

$$Avg = \frac{\sum_{i=1}^{k} \frac{B_i}{A_i}}{k} \text{ X } 100\% \qquad (2)$$

where, *k* is the number of sentences pattern, *B* is the number of parse tree correctly proposed, and *A* is the number of related pattern.

$$WAvg = \frac{\sum_{i=1}^{k} B_i}{N} \text{ X } 100\% \qquad (3)$$

where, *k* is the number of sentences pattern, *B* is the number of parse tree correctly proposed, and *N* is the total sentences in testing dataset.
        The evaluation result is shown in Table 2.

TABLE 2. Evaluation result

Sentence Pattern	No. of test case	No. of correct tree	% Correct proposed parse tree
FN + FN	28	25	89.30%
FN + FK	69	67	97.10%
FN + FA	40	40	100%
FN + FS	33	33	100%
		Average	96.60%
		Weighted Average	97.10%

The evaluation result shows that the parser prototype is able to achieve weighted average rate of 97.1%. This higher score recognizes that the bottom-up technique, with support of statistical approach, is able to minimize ambiguity and parse Malay language sentences effectively. The highest results are achieved by FN + FA and FN + FS patterns, while the FN + FK and FN + FN have achieved 97.1% and 89.3% respectively. Two major factors which lead to low FN + FK and FN + FN are ambiguity in selecting the appropriate grammar rules and lexical ambiguity, which create confusion in applying appropriate grammar rules and part of speech to represent the sentence. The results show that the Malay Parser is able to suggest the most appropriate parse tree based on the probability value of words matched with the grammar rules. The number of test cases that incorrectly proposed the most precise parse trees have been revisited. Based on our finding, the main factors that influence the generating and proposing the parse tree fall under two areas; the same probabilistic value for words that have more than a POS and sentence syntactic structure that can represent more than a pattern. Therefore, a larger number of training dataset is predicted to improve the results. In general, the Malay Parser can be used to solve ambiguity problem in syntactic ambiguity in Malay language analysis.

CONCLUSION

The developed parser is able to derive parse trees for Malay language sentences. An evaluation of the parser performance shows its ability to propose the most precise parse tree if a sentence syntactic structure produces more than a parse tree. Meanwhile, the effectiveness of bottom-up parsing technique, supported with statistical approach in Malay language sentences, is acceptable. The parser is able to solve syntactic ambiguity in a simple Malay sentence by proposing the most appropriate parse tree based on probability calculation and grammar rules. The parser can be extended to measure its effectiveness in other languages especially indigenous languages (Kadazan, Dusun, etc), provided that the languages syntactic structure is in CFG form and sufficient datasets are collected for training and testing data.

In general, solving the ambiguity will help researchers to acquire insight information on the sentence structure which is a crucial input to support complex NLP tasks such as semantic and pragmatic analysis. However, more efforts will be needed to analyze a complex sentence structure. The dataset could be improved by adding complex sentences from different sources either from well-structured texts (books, articles etc.) or unstructured texts (social media) and also classical Malay texts. Analyzing such texts will be challenging as more grammar rules will be needed.

REFERENCES

Ahmad I. Z. Abidin, S.P. Yong, Rozana Kasbon & Hazreen Ahmad. (2007). Utilizing Top-Down Parsing Technique in the Development of a Malay Language Sentence Parser, Proceeding of the 2nd International Conference on Informatics. pp. 125-131.

Alavi, M. and, Umanath, N.S. (1989). Application Software Prototyping, in Encyclopedia of Computer Science and Technology. Vol. 21(6). In Kent, A. & Williams, J. G. (Eds.). *ADA and Distributed Systems to Visual Languages* (pp. 19- 38). Florida: CRC Press.

Boar, B.H. (1984). *Application Prototyping: A Requirements Definition Strategy of the 80s*: New York: John Wiley & Sons,.

Charniak, E. (1993). *Statistical Language Learning*. Cambridge: MIT Press.

Charniak, E. (1997). Statististical Techniques for Natural Language Parsing. *AI Magazine. Vol. 18*, 33-43.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Collins, M.J. (2003). Head-Driven Statistical Models For Natural Language Parsing. *MIT Press Journal. Vol. 29* (4), 589-637.

Dale, R. (2000). Symbolic Approaches to Natural Language Processing. In Dale, R., Moisl, H. and Somers, H. (Eds.), *Handbook of Natural Language Processing* (pp.1-9). New York: Marcel Dekker.

Hashim Musa. (1990). *Sintaksis Bahasa Melayu: Suatu Huraiann Berdasarkan Rumus Struktur Frasa*. Kuala Lumpur: Agensi Penerbitan Nusantara.

Hassan Mohamed, Nazlia Omar & Mohd Juzaiddin. (2011). Malay Part of Speech Tagger, A comparative study on Tagging Tools. *Asia-Pacific Journal of Information Technology and Multimedia. Vol. 4*(1), 11-23.

Hassan Mohamed, Nazlia Omar & Mohd Juzaiddin. (2011). Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach. International Conference on Semantic Technology and Information Retrieval. pp.231-236. Putrajaya.

Hopcroft, J.E., Motwani, R. & Ullman., J.D. (2006). *Introduction to Automata Theory, Languages, and Computation*. Boston: Pearson Addison Wesley.

Jurafsky, D. & Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice-Hall.

Knowles, G. O. & Zuraidah Mohd Don. (2006). *Word Class in Malay: A Corpus-Based Approach*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa & Abdul Hamid Mahmood. (2008). *Tatabahasa Dewan*, *3rd ed*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nik Safiah Karim. (1975). The Major Syntactic Structures of Bahasa Malaysia and their Implication of the standardization of the Language. Unpublish PhD Thesis, Ohio University.

Noor Hafhizah Abd Rahim. (2011). A Statistical Parser To Reduce Structural Ambiguity in Malay Grammar Rules. Unpublish Master Thesis, Malayan University of Malaysia.

Nur Ashikin Halid, Nazlia Omar. (2017). Malay Part Of Speech Tagging Using Ruled-Based Approach. *Asia-Pacific Journal of Information Technology and Multimedia*. *Vol. 6*(2).

Palmer, D.D. (2000). Tokenisation and Sentence Segmentation. in R. Dale, H. Moisl and. H. Somers(Eds.). *Handbook of Natural Language Processing*. New York: Marcel Dekker, pp. 11-35.

Resnik, P. & Lin, J. (2013). *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley Blackwell.

Rozana Kasbon, Nurul Atiqah Amran, Eliza Mazmee Mazlan & Saipunidzam Mahamad. (2011). Malay Language Sentence Checker. *World Applied Sciences Journal*. *Vol. 12*, 19-25.

Sabrina Tiun, Rosni Abdullah, Tang Enya Kong, Siti Khaotijah Muhammad. (2011). Korpus Pertuturan Sintaksis-Prosodi Bahasa Melayu. *Asia-Pacific Journal of Information Technology and Multimedia. Vol. 2*(1), 1-12.

Yusnita Mohd Noor & Zulaikha Jamaluddin. (2012.) Malay declarative Sentence: Visualization & Sentence Correction. IEEE Conference on Open Systems. pp. 1-5.

Zulkifley Hamid, Ramli Md Salleh & Rahim Aman. (2015). *Linguistik Melayu*. Bangi: Penerbit Universti Kebangsaan Malaysia.

ABOUT THE AUTHORS

Muhammad Azhar Fairuzz Hiloh is a former master student at the Center of Artificial Intelligence Technology (CAIT) at the Faculty of Information Science & Technology, National University of Malaysia. His main area of interest is the study of computational linguistics and his main research area is syntax analysis in Bahasa Melayu, which he formalized in his Master degree.

Prof. Dr. Mohd Juzaiddin Ab Aziz is a lecturer in the Research Center for Software Technology & Management at the Faculty of Information Science & Technology, National University of Malaysia. His area of interest and research is the Natural Language Processing, Machine Translation and Text Classification. He is actively involves in multiple research grants and currently he is the Director of Center For Information Technology, UKM.

Dr. Lailatul Qadri Zakaria is a senior lecturer in the Center of Artificial Intelligence Technology (CAIT) at the Faculty of Information Science & Technology, National University of Malaysia. She did her PhD at Southampton University in 2011. Her area of interest and research are in the Natural Language Processing, Semantic Web and Ontology and currently, she is the head of lab for Asian Natural Language Processing in CAIT.