Identification of Ambiguous Sentence Pattern in Indonesian Using Shift-Reduce Parsing

M. Fachrurrozi[1], Novi Yusliani[2], Muharromi Maya Agustin[3]

Informatics Engineering Department, Faculty of Computer Science, Sriwijaya University,

[1]**Obetsobets@gmail.com**, [2]**novi.yusliani@gmail.com**, [3]**hayastin@gmail.com**

Abstract— In a sentence writing, mistakes often happened unconsciously by the writers that caused a sentence has double meaning. This research developed a software that can identify the pattern of ambiguous sentences in Bahasa using Shift Reduce Parsing. Input for this software is a single sentence in Bahasa. This research involved three preprocessing namely case folding, tokenizing, and Part of Speech (POS) Tagging. Then the result of preprocessing used in identifying ambiguous sentence. In the process of identifying ambiguous sentences, Shift Reduce Parsing used to the parsing's result of a sentence. As the result, this software that can differentiate whether if it is an ambiguous sentence or not. Accuracy of identifying the pattern of ambiguous sentence in Bahasa using Shift Reduce Parsing reaches 83%.

Keywords— Ambiguous Sentense, Shift Reduce Parsing, Natural Language Processing.

I. INTRODUCTION

Sentences that have more than one possible meaning of a sentence is called ambiguous sentence. Ambiguous sentences can hamper communication because the receiver (listener or reader) can interpret different meanings with the meaning intended by the speaker or writer[1]. This can happen either because the structures or patterns due to the use of his or her sentence. Humans can distinguish clearly the error occurred but not with the computer, the computer can not identify a particular sentence patterns.

One method that can be used in a sentence is to recognize patterns shift reduce parsing methods. There are two processes that occur in this method, namely: shift and reduce operations. Shift operation is the addition operation of said input sentence on top of the stack elements (piles) which is referred to as the top. While surgery is the surgical removal reduce top element on the stack and replace it with a new element in the form of a grammar rule corresponding elements are replaced such information. This method is often used in terms of grammatical categorization. This method will split the input sentence into words in the form of a stack. After that, some of these words can be recognized identity so that he will be processed in a particular grammar[2].

II. AMBIGUOUS SENTENCE

Sentences that have more than one possible meaning of a sentence is called ambiguous sentence. Ambiguous sentences can hamper communication because the receiver (listener or reader) can interpret different meanings with the meaning intended by the speaker or writer [3]. Examples of ambiguous sentences can be seen in Table II-2.

Table 1. Examples of Ambiguous Sentence

No.	Kalimat
1.	Tahun ini gaji karyawan baru dinaikkan.
2.	Tahun ini gaji karyawan-baru dinaikkan.
3.	Gaji karyawan tahun ini baru dinaikkan.

The word 'baru' in the sentence (1) in Table II-2 can lead to more than one interpretation (ambiguous). If the word 'baru' employee explained, the sentence can use hyphens and turned into sentences (2). If the word 'baru' to explain the increased (previous salary never increased), the sentence was commuted to a sentence (3).

III. METHODOLOGY

The software accepts input in the form of a single sentence in Indonesian which will then be checked using the sentence pattern Shift-Reduce Parsing. The resulting output is the truth value obtained from the shift and reduce. If at the end of the process reduce accepted and

have more than one parsing process, then input the correct sentence for an ambiguous sentence patterns. Software architecture to be built can be seen in Figure 1.
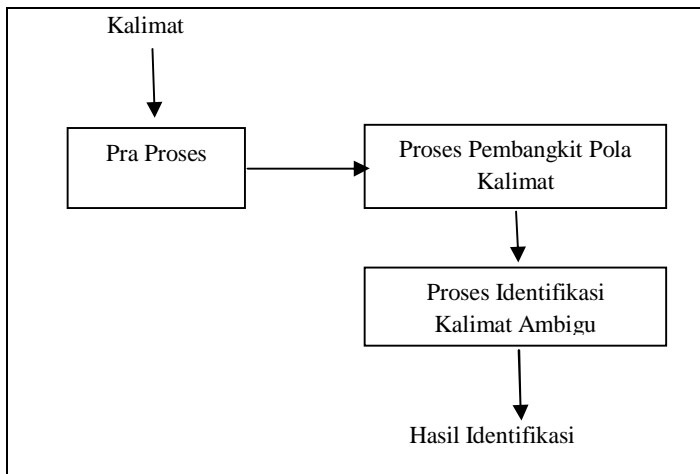

Figure 1. Software Architecture

A. Preprocessing

Preprocessing is beginning the process of managing data before data processing is done[4]. The purpose of this stage is to remove the characters other than letters and break the sentence into a collection of words. Stages including preprocessing is performed on case folding and tokenizing. Case folding is changing all the letters in the document to lowercase. Only the letter 'a' to 'z' are acceptable [5]. Considered a delimiter character other than letters.

B. Sentence Pattern Recognition

This process is a stage that serves to determine the identity of the words of each word in the input sentence. Steps being taken in this process are Part of Speech (POS) Tagging and generalization. POS tagging is a process that is done to determine the type of a word in the text. Type the word in question is a grammatical category, ie verbs, nouns, adjectives, adverbs, and so on. POS tagger is an important tool in many natural language processing applications such as word sense disambiguation, decomposers (parser), question answering, and machine. Input of a POS Tagger is a sentence and output in the form of tags for each word contained in the input sentence.

Hidden Markov Model is a statistical model where the system being modeled is assumed to be a Markov process with unobservable conditions. Hidden Markov models have a hidden condition (hidden states), the value of output (observation), the transition probability, possibility of emissions, and an initial condition.

Tag represents the hidden states represent the state and observation. Transition probabilities depend on the state of his tag partner. Emission probabilities depend on the applicable tag.

POS Tagger based on hidden Markov models has a running time better than the other such as the probabilistic method developed by Wicaksono dan Purwarianti[6].

C. Identification Sentence

This process is a stage that serves to determine the pattern of the sentence is ambiguous or not of an input sentence. The method used to determine the pattern of the sentence is ambiguous or not the Shift-Reduce Parsing.

The process of parsing using Shift-Reduce Parsing can be seen in the parsing tree as follows:
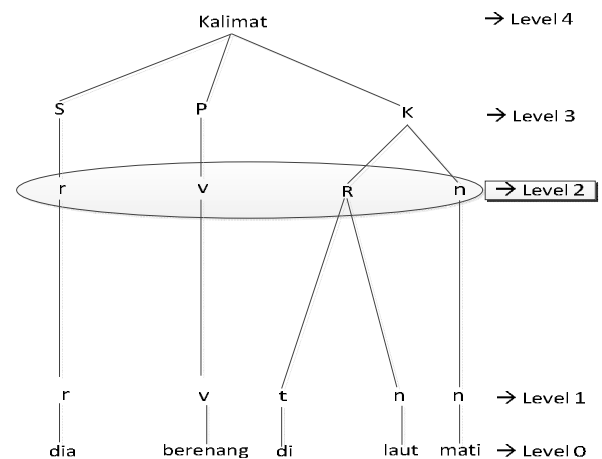

Figure 2. Type-1 Parsing Sentence

Figure 2. shown the parsing process type-1 using Shift Reduce in sentence "Dia berenang di Laut Mati".
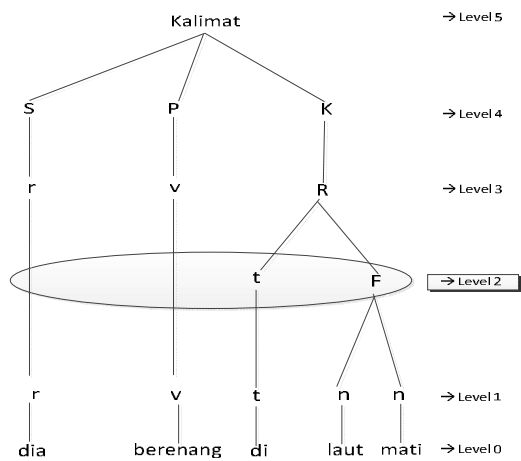
Figure 3. Type-2 Parsing Sentence

In the description of the type-1 (Figure 2), the meaning of the sentence is that he swam in the ocean, where the sea is the name of the dead sea. While the results to the description of the type-2 (Figure 3), the ambiguous sentence can be interpreted as follows: he was swimming in the sea, and then he died. From the above, it can be concluded that the sentence that has more than one meaning called ambiguous sentence.

The process of determining ambiguous sentences can be seen through the tree parsing, where there is more than one type of tree parsing of one sentence. Parsing process changes can be seen in the level 2 tree parsing in Figure 2 and Figure 3. At level 2 type-1 to the parsing, resulting in the sentence pattern is r + v + R + n, while the 2nd type of parsing the resulting pattern is r + v + t + F.

IV. SHIFT REDUCE PARSING

Shift Reduce Parsing (SR Parsing) is a parsing technique which belongs to the category of bottom-up parsing. SR parsing used as tokens and form a production line to build the parse tree (parse tree). SR Parsing using stack in order to maintain the order of each token. SR Parsing is done after case folding process, tokenizing, and Part of Speech (POS) Tagging, is used to determine the sentence pattern of a sentence. In general, the Shift-Reduce Parsing algorithm has the following action:

1. Shift, add one element (token obtained from the input) on the stack. Shift action is only the displacement (shifting) the first item (the top of the pile of words, in this case the per-item(form a single word) of the RHS (Right Handle Stack) to the LHS (Left Handle Stack)

2. Reduce, remove top element on the LHS and replace it by adding one element corresponding nonterminal.

If the LHS is empty, then the only action that can dilakukan. If RHS Shift empty, Reduce the only action performed. If the RHS and LHS is not empty, then there is a possibility of that happening is both action, and processors must provide the conditions for menenentukan action taken. If the action taken is reduce, then determined a non-terminal (in this rule) what should be added to the top of the LHS LHS replace the item itself. If the action is performed Shift, it will form a new terminal as the leaf nodes of the parse tree and will form a new subtree. In software design, the input sentence in the form of sentences in Indonesian then the syntax rules of syntax used is Indonesian.

V. IMPLEMENTATION AND RESULT ANALYSIS

The system implemented ambiguous sentence pattern identification is done by inserting a sentence in FormUtama. The number of sentences which tested 30 samples of sentences, where each sentence to be tested is a single sentence, in the form of ambiguous sentences and the sentence is not ambiguous.

Testing is done by comparing the amount of data input sentence successfully tested and the overall amount of data input sentences tested. The accuracy of the software has reached 83.33%. Based on the results of testing of 30 samples of sentences, obtained 5 sentences can not be checked with the appropriate sentence patterns. Errors caused by several things, such as:

1. Tagging error occured when the defined word class process.
2. The input sentence does not belong to a single sentence patterns are usually the subject is in the beginning of the sentence.

VI. Conclusion

The conclusion that can be obtained from this research are : Shift Reduce Parsing can be used to identify patterns of sentences in Indonesian and can be applied to the case of Ambiguous Sentence. The accuracy of identification of ambiguous sentence pattern in Indonesian using Shift Reduce Parsing method on software built by 83.3 %, and results of pattern recognition ambiguous sentence is determined by the defined word class, word class only of each word in a

phrase that is used to process the Shift Reduce Parsing. Therefore , the software can not recognize an ambiguous sentence in appropriate if there is an error in the definition of the word class process.

REFERENCES

[1] Wiyanto,A. 2012. *Kitab Bahasa Indonesia*. Yogyakarta: Jogja Bangkit

Publisher.

[2] Shieber, S. M. 2002. Sentence Disambiguation by a Shift-Reduce Parsing Tecnique.

[3] Wiyanto,A. 2012. *Kitab Bahasa Indonesia*. Yogyakarta: Jogja Bangkit Publisher.

[4] Hatta, A., Ramadijanti, N., & Helen, A. 2010. Rancang Bangun Sistem Pengelolaan Dokumen-dokumen Penting Menggunakan Text Mining. Surabaya: Institut Teknologi Sepuluh Nopember.

[5] Triawati, C. 2009. *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi.* Bandung: IT TELKOM.

[6] Wicaksono, A. F., & Purwarianti, A. 2010. HMM Based Part-of-Speech Tagger for Bahasa Indonesia. *Proceeding The Forth International MALINDO Workshop.* Jakarta, Indonesia.