Indonesian-English Machine Translation Using Rule-Based Method

[1]Novi Yusliani, [2]Yunita, [3]Wenty Octaviani
[1,2,3]Sriwijaya University
novi_yusliani@unsri.ac.id, yunita.v1t4@gmail.com, wentyoct@gmail.com

***Abstract*-, Rule-Based Machine Translation (RBMT) used a set of linguistic information to translate source language to target language. POS tagger and Shift-Reduce-Parsing (SRP) could be used to get the linguistic information. POS tagger was used to get word class of each word in sentence and SRP was used to get the function of each word in sentence. SRP was also used to get the structure of sentence. In this research, POS tagger and SRP were used to get the linguistic information of source sentence. Translation process was done by using billingual dictionary. Last, a set of rules was used to generate the target sentence. The accuracy of Indonesian-English machine translation was 100% for the S-P-Adv pattern, but for the S-P pattern and S-P-O pattern is 93,33%.**

BACKGROUND

English is an international language. This language is standard communication language between people from different country. As a consequence, many text documents in internet are using English. In order to get a better understanding about the text, people will translate the text into their native language. Translating English text manually is not easy for them who do not understand English. It is difficult for them to understand English text. Usually, they will search text using their native language.

Machine translation (MT) is subfield of computational linguistic which is want to make a computer can translating text or speech from source language to the target language while preserving the meaning and interpretation [4][6][9]. MT is trying to help people who do not understand with the source language. Beside that, MT is also trying to help people who want to learn translating the source language into their native language. Since natural languages are highly complex. Developing MT needs an understanding with the both of the natural language, source language and target language. Every natural language has different grammar.

Rule-Based Machine Translation (RBMT) is using rule-based approach to translate the source language to the target language. RBMT needs a linguistic information of the source language and the target language. In RBMT, linguistic information is used to get linguistic rule between both of the language [7]. There are three processes in RBMT, morphological process, syntax analysis, and transformation process (structural and lexical) [1]. Morphological process and syntax analysis are aimed to get linguistic information of a language. Then, the transformation process are aimed to translate the source language into the target language.

In the previous research, Indonesian-English RBMT was developed by using Apertium platform, which contains three modules [1]. The modules are morphological analyzer, morphological disambiguator, and lexical transfer and structural transfer. The morphological analyzer module used Indonesian dictionary and Indonesian morphology analyzer web service. The morphological disambiguator module used rule based morphological disambiguation by using Constraint Grammar tools. Then, the lexical and structural transfer module worked based on bilingual dictionary. This research attempts to build another RBMT for Indonesian to English using HMM based POS Tagger [2], Shift-Reduce-Parsing [3][5], and billingual dictionary.

RESULT AND DISCUSSION

An Indonesian-English machine translation using rule-based method of this research has four modules, namely, POS Tagger, structure identification, translating word, and sentence generator. POS Tagger module used HMM based POS Tagger [2]. This module is used to analyze the input sentence. Output from this module is word class of each word in the sentence. Structure identification module used shift-reduce parsing. This module was used to analyze the function of each word in the sentence. Translating word module used to translate each word from the source language to the target language. This module used billingual dictionary (Indonesian-English) to translate. Last module is sentence generator. It used to generate the sentence translation. This module used a set of rule to generate the sentence. The architecture of Indonesian-English machine translation using rule-based method can be seen in Fig 1.

First, input sentence will be processed in POS Tagger module. In this module, there are three processes. The first process is making all characters in the sentence into

ISBN: 979-587-573-6

small character. Second process is to tokenize the sentence into individual word. Last process is identifying word class of each word in the input sentence. All processes in POS Tagger module can be seen in Table 1. Word class identification is using Hidden Markov Model (HMM) based POS Tagger library [2]. If word class of each word in the input sentence has already known, then the input sentence will be processed in structure identification module.

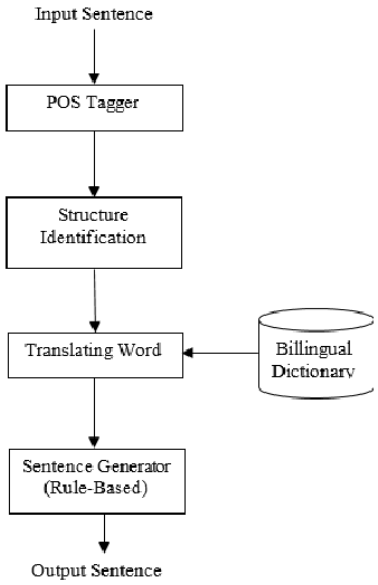

Fig 1. Architecture of Indonesian-English Machine Translation using Rule-Based Method

Table 1. POS Tagger Module

Input Sentence	: Saya memotong rumput "I cut the grass"
First Process	: saya memotong rumput
Second Process	: saya memotong rumput
Third Process	: saya (Pronoun) memotong (Verb) rumput (Noun)

Structure identification module is used to analyze the sentence structure. If the structure was not matched with the pattern of Indonesia sentence, then the sentence could not be processed. In this research, patterns of Indonesia sentence could be processed were Subject-Predicate (S-P), Subject-Predicate-Object (S-P-O), and Subject-Predicate-Adverb (S-P-Adv) [8]. This module was also used to identify noun-phrase in the input sentence. Output from this module was function of each word in the sentence. Fig 2 shows the process in structure identification module.

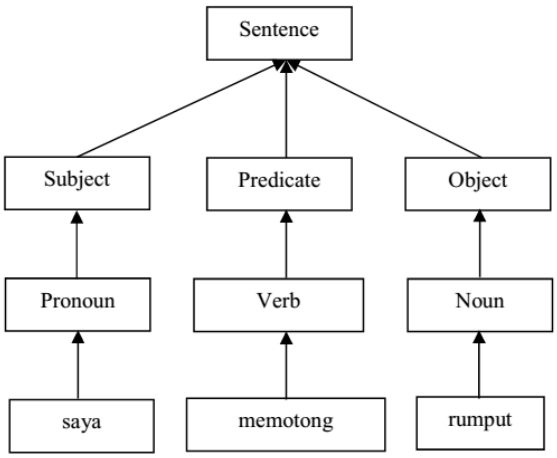

Fig 2. Structure Identification Module

Table 2. Translating Word

Sentence : saya memotong rumput		
Word	Function	Translation
saya	Subject	i
memotong	Predicate	cut
rumput	Object	grass

Table 3. Rules to Generate Target Sentence

Rules to generate target sentence, are :

1. If subjects are They, We, You, or I then after subject is verb 1 form
   Example :
   Mereka berlari "They run"
2. If subjects are He or She then after subject is verb 1 + s/es form
   Example :
   Dia tertawa "He laughs"
3. If the subject of the input sentence is using name then the subject will not be translated
4. If subjects are He or She and after the subject is adverb then subject + 'is'
   Contoh :
   Dia tampan "He is handsome"
5. If subject is I and after the subject is adverb then subject + 'am'
   Contoh :
   Saya cantik "I am pretty"
6. If subjects are They, We, or You and after the subject is adverb then subject + 'are'
   Contoh :
   Mereka pintar "They are smart"

Translating word module is used to translate each word in the source language to the target language. Translation could be done by using billingual dictionary (Indonesian-English). Table 2 shows the process in translating word

module. If word could not be translated into English, it might be the word was not in the dictionary. Sentence generator module is used to generate the target sentence. Sentence could be generated by seeing the function of each word in the input sentence and using rules. Table 3 shows rules to generate the target sentence (English sentence). If the subject of the target sentence using "He" or "She" then the verb would be translated into verb 1 with s/es form by checking in the verb dictionary. Verb dictionary consists of a set verb 1 form with its verb 1 plus s/es form. Pattern of the target sentence was composed based on the English grammar. In this research, sentence could be translated into English is Indonesia sentence using present time.

Table 4. Experiment Result

Pattern	Sentence	Correct Translation	Incorrect Translation
S-P	15	14	1
S-P-O	15	14	1
S-P-Adv	15	15	0

In order to test the accuracy of the Indonesian-English machine translation using rule-based method, 45 sample sentences has been tested. Sample sentences consist of 15 sentence with S-P pattern, sentence with S-P-O pattern, and 15 sentences with S-P-Adv pattern. Each of sample sentences consist of Indonesian sentence and its English sentence translation. Sample sentences were taken from book and internet which had already been checked by the expert. Result of these experiments can be seen on Table 4. The accuracy of this machine translation for the S-P-Adv pattern are 100%, but for the S-P pattern and S-P-O pattern, the accuracy is 93,33%. Incorrect translation is caused by the word translation in dictionary and in sample is different. Incorrect translation is also caused by the lack of machine translation in processing word which has possessive sufix.

REFERENCES

[1] E. Yulianti, I. Budi, A. N. Hidayanto, H. M. Manurung, and M. Adriani, "Developing Indonesian-English Hybrid Machine Translation System, " in *The 3rd International Conference on Advanced Computer Science and Information Systems (ICACSI)*, 2011.

[2] A. F. Wicaksono dan A. Purwarianti, "HMM based POS Tagger for Bahasa Indonesia," Proceeding of the Fourth International MALINDO Workshop (MALIND0 2010), Agustus 2010.

[3] M. Fachrurrozi, N. Yusliani, and M. M. Agustin, Identification of Ambigous Sentence Pattern in Indonesian using Shift Reduce Parsing. *International Conference on Computer Science and Engineering*, Oktober 2014.

[4] C. Kit, H. Pan, dan J. Webster, Example-Based Machine Translation: A New Paradigm, Translation and Information Technology, Chinese U of HK Press, 2002, pp. 57-78.

[5] S. M. Shieber, Sentence Disambiguation by a Shift-Reduce Parsing Technique, *In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, 1983, pp. 113-118.

[6] M. R. Costa-Jussa, M. Farrus, J. B. Marino, and J. A. R. Fonollosa, Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems, *Computing and Informatics*, Vol. 31, 2012, pp. 245-270.

[7] S. Tripathi, and J. K. Sarkhel, Approaches to Machine Translation, *Annals of Library and Information Studies*, Vol. 57, 2010, pp. 388-393.

[8] H. Alwi, S. Dardjowidjojo, H. Lapoliva, and A. M. Moeliono, Tata Bahasa Baku Bahasa Indonesia, Balai Pustaka, 2003.

[9] T. D. Singh, and B. Sivaji, Manipuri-English Example Based Machine Translation System. *International Journal of Computational Linguistics and Applications (IJCLA)* Vol. 1, No. 1-2, Jan-Dec 2010.

ISBN: 979-587-573-6