Simplification Complex Sentences in Indonesia Language using Rule-Based Reasoning

Rifka Widyastuti[1], M. Fachrurrozi[2], Novi Yusliani[3]
Informatics Bilingual Engineering, Faculty of Computer Science, Sriwijaya University
Palembang, Indonesia
[1]rifka.widyastuti@gmail.com
[2]Obetsobets@gmail.com
[3]novi.yusliani@gmail.com

***Abstract*-Complex sentence consist of two or more single sentence and one or more conjunctions. Complex sentences have many meaning, so that is quite difficult to be simplified because the results of simplification complex sentence does not match the original sentence. To solve this problem, this research using Rule-Based Reasoning method and Surface Expression Rules. Because Rule-Based Reasoning method can be adapted to the rules of surface expression that can look for patterns of complex sentences, so thus simplifying precise and accurate results. The results of researchis the Rule-Based Reasoning methods used in accordance with the accuracy of 93.8% and an assessment of questionnaires obtained values 76-100.**

***Keyword*–complex sentences, simplification, rule-based reasoning, surface expression, pattern complex sentences.**

I. INTRODUCTION

Complex sentence consist two or more single sentence. Every complex sentences has different conjunctions and some are not wearing but use a comma.Complex sentences can be determined by looking at the use of conjunctions or punctuation used. People often difficult to understand what you want delivered in complex sentences. Using conjunctive or sign can make a complex sentence into a different meaning and content of information.

The problem is when simplification the sentence which divides complex sentences into a single sentence in which a single sentence that has been simplified to alter the meaning and content of information. Simplifying text (text simplification) is one of the fields of natural language processing (NLP) which rewrite a sentence to reduce syntactic complexity (syntactic complexity) and lexical complexity (lexical complexity) without changing or eliminating the meaning of the sentence and fill in the information sentence[1]. In particular, the simplification of the sentence has been developed in various countries. Development is done by a variety of methods and rules. However, in Indonesia the development of sentence simplification is not much, especially simplification complex sentences

II. RELATED WORK

Some research on text simplification done. In this study, will try simplifying the text in Indonesian to simplify complex sentences. Previous research [2] studied the question and answer system development for Non-factiod question for Indonesian. Non-factoid question is a question-answer questions that generally require a fairly lengthy explanation as a definition of a term method for doing something or the cause of an incident. This indicates that non-factoid question is more complex, because it requires the classification of questions to get the expected answer. Wear pattern classification question classification based on pattern Surface Expression. Types of questions can be included to predict the type of response generated.

Future studies related on the simplification of text in French[3] entitled Acquisition of Syntactic Simplification Rules for French. His research describes the simplification of syntax (syntactic simplification) is a data-driven approach that implements two methods. The first method is manual corpus analysis that aims to identify the word you would be simplified, then the second method is a semi-automatic that automatically identifies the simplified function informs sentence simplification rules. The results of his research to overcome obstacles no longer need the data as parallel resources and increase flexibility. In particular, syntactic simplification can explore domains on user-generated content as pre-editing for statistical machine translation.

Research simplification of text above, this study will examine the simplification of complex sentences in Indonesian using Rule-Based Reasoning method based on rules and their surface expression tagger post on the introduction of a class of words in a sentence. In a complex sentence simplification is expected to facilitate the delivery of simplifying complex sentences and the meaning of the sentence.

III. PERFORMANCE

Natural Language Processing is the area of research and application that explores how computers can be used to understand and manipulate the natural language text or speech to do something useful [4].

A Preprocessing

Preprocessing is process of managing the data before the processing data [5]. Preprocessing consist of case folding and tokenizing. Case folding is process of changing all the letters in a document / sentence to lowercase. Only the letters 'a' through 'z' received [6] while the characters other than letters received are considered delimiter. Examples delimiter can be seen in Table I.

TABLE I
DAFTAR *DELIMITER*

Daftar *Delimiter*									
0	5	[	%	`	.	?	|	)	≥
1	6	]	^	~	,	:	!	-	∞
2	7	{	&	\\	/	;	@	_	π
3	8	}	*	£	<	‘	#	+	±
4	9	\	(	€	>	‘’	$	=	ϕ

Tokenizing is process of identification the smallest units (tokens) of a sentence structure (Triawati, 2009). Breaking sentences into single words performed by scanning a sentence using white space separators such as spaces, tabs, and newline. Schematic of the process of folding and tokenizing case can be seen in Table II.

TABLE II
PREPROCESSING SENTENCES SCHEME

Preprocessing Sentences	
Sentences :	Ibu Pergi Ke Pasar
Case folding :	ibu pergi ke pasar
Tokenizing :	"ibu" "pergi" "ke" "pasar"

B Part of Speech

Part of Speech (POS) tagging is a process that is done to determine the type of a word in the text. A simple form of this process is the identification of words as adjectives, adverbial, interjection, conjunction, noun, numerial, prepositions, pronouns, verbs, etc. [5]. The process of determining the type of words in a sentence can be seen in Figure 1.

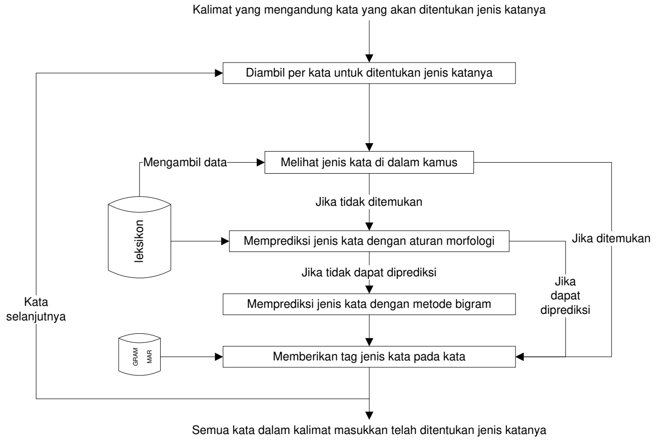

Kalimat yang mengandung kata yang akan ditentukan jenis katanya
Diambil per kata untuk ditentukan jenis katanya
Mengambil data
Melihat jenis kata di dalam kamus
Jika tidak ditemukan
leksikon
Memprediksi jenis kata dengan aturan morfologi
Jika ditemukan
Jika tidak dapat diprediksi
Jika dapat diprediksi
Kata selanjutnya
Memprediksi jenis kata dengan metode bigram
GRAM MAR
Memberikan tag jenis kata pada kata
Semua kata dalam kalimat masukkan telah ditentukan jenis katanya

Fig. 1 Process of Identification Type of Word[5]

C Rule-Based Reasoning

Rule-Based Reasoning is a decision support system which also has a knowledge base. In this method, the settlement of the problem based on an artificial intelligence approach using problem-solving techniques based on the rules contained in the knowledge base [7].

[2] uses the rules of the component surface expression answer finder. Surface expression is the surface expression of the sentence or the pattern used in the sentence. regulation of surface expression in the study can be found in appendix

D Complex Sentences

Complex sentence is a merger of two or more single sentences using conjunctions. Examples of complex sentence simplification can be seen in Figure 2.

1. Tini berbelanja sayuran.
2. Tini memasak sayuran
3. Tini berbelanja sayuran dan memasaknya

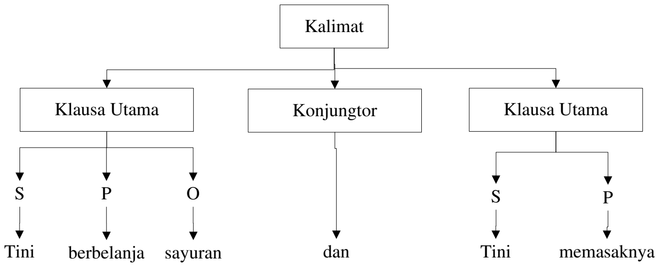
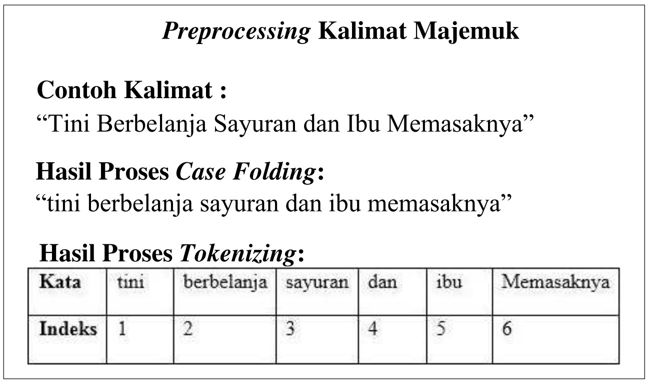


Fig. 2 Chart of Complex Sentences (3)

IV. EXPERIMENTAL

Simplification complex sentence is not as easy as one might imagine, some people find it difficult to simplification complex sentences, especially during the learning process in schools. Therefore, need a applications to help the learning process and make it more attractive. In this research, simplification complex sentence process starts from preprocessing which case folding and tokenizing. The results of the research complex sentence preprocessing can be seen in Figure 3.



***Preprocessing* Kalimat Majemuk**

Contoh Kalimat :
"Tini Berbelanja Sayuran dan Ibu Memasaknya"

Hasil Proses *Case Folding*:
"tini berbelanja sayuran dan ibu memasaknya"

Hasil Proses *Tokenizing*:

Kata	tini	berbelanja	sayuran	dan	ibu	Memasaknya
Indeks	1	2	3	4	5	6

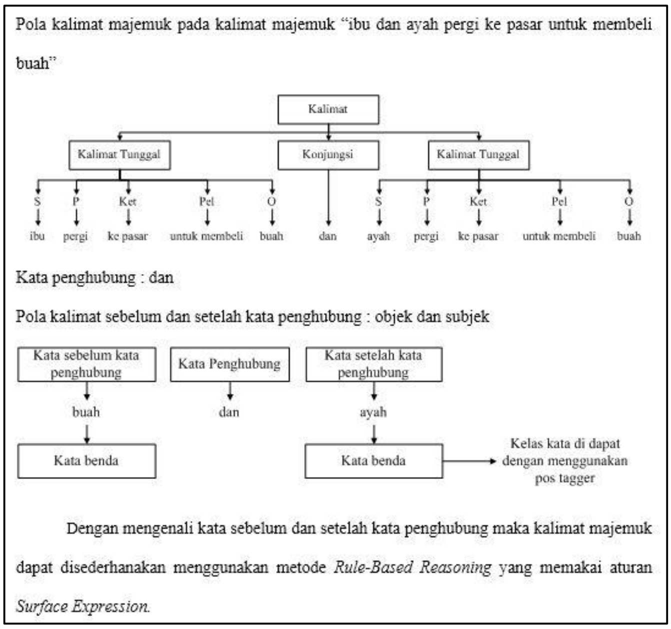
Fig. 3 Preprocessing of Simplification Complex Sentences

After we get a results in the form of preprocessing tokens (word class), tokenizing on this journal wear NLP_ITB package where the package is Indonesian word dictionary. Token can make easy to process of simplification of complex sentences. Further, simplification complex sentences of the process that is using the Rule-Based Reasoning with rules Surface Expression. Process simplification of complex sentences can be seen in Figure 4 and Figure 5.

This research used 60 samples were taken from the complex sentence http://bse.kemdikbud.go.id/. Based on the experiment results of the software by entering the 60 samples of complex sentences, obtained 4 sample of complex sentences that can not be simplified accurately. Experiment result on this research using 60 sample of complex sentences can be seen on appendix B. This is due to several factors that the sentences can not simplified accurately, there are:

1. The token tagging errors occurred in the compound sentence "his face is thin and pale". The error occurs on the token marking words that should generate "n, v, c, n, v", but in POSTagging generated token is "v, n, c, v, n". The error occurs from the package NLP_ITB.

2. Sentence Compound "It's fun playing ball so spaced out" can not be reduced to a single sentence and a two conjunctions appropriately. The fault lies in the pattern of compound sentence has no subject. So that can not be simplified complex sentences correctly.
3. Compound sentence "the birthday party would not be more festive if you come to attend". The error occurs on the token marking the word "if" is a word that should be connecting, but the token POSTagging recognizable words with a noun. The error occurs from the package NLP_ITB.
4. Sentence compound "People panic because there was an earthquake" in POSTagging identified by the token of the word "n, n, c, v, v", but should have obtained a token word is "n, v, c, v, n". The error occurs from the package NLP_ITB.


Pola kalimat majemuk pada kalimat majemuk "ibu dan ayah pergi ke pasar untuk membeli buah"

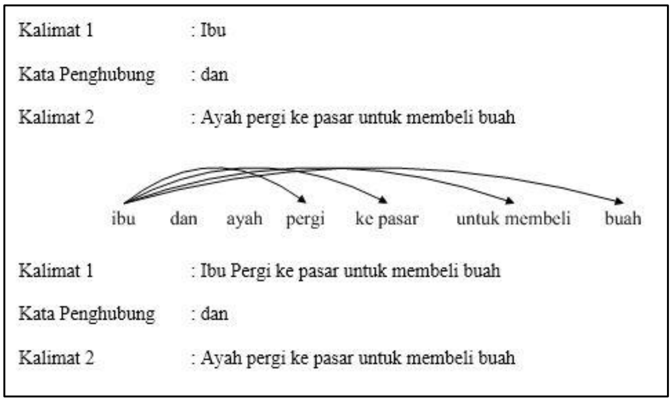
Kata penghubung : dan

Pola kalimat sebelum dan setelah kata penghubung : objek dan subjek

Kata sebelum kata penghubung
Kata Penghubung
Kata setelah kata penghubung
buah
dan
ayah
Kata benda
Kata benda
Kelas kata di dapat dengan menggunakan pos tagger

Dengan mengenali kata sebelum dan setelah kata penghubung maka kalimat majemuk dapat disederhanakan menggunakan metode *Rule-Based Reasoning* yang memakai aturan *Surface Expression.*

Fig. 4 Surface Expression Rules in Simplification Complex Sentences


Kalimat 1 : Ibu

Kata Penghubung : dan

Kalimat 2 : Ayah pergi ke pasar untuk membeli buah

Kalimat 1 : Ibu Pergi ke pasar untuk membeli buah

Kata Penghubung : dan

Kalimat 2 : Ayah pergi ke pasar untuk membeli buah

Fig. 5 Simplification Complex Sentences using Rule-Based Reasoning

Based on the experimental results of 60 samples of complex sentences obtained 4 sample of complex sentences that can not be simplified appropriately. Therefore, the percentage of success of software obtained for 93.3% of the software is built. Word tokens not generated as expected. However, the word is sometimes different tokens if put in a different sentence. Therefore, there are some words that can not be replaced tokens he said. For example the word "if" is not a word said base so that the resulting token is different.

V. CONCLUSION

The conclusion that can be take from this study are

1. Methods of Rule-Based Reasoning can be used to simplification sentence and can be applied to the case of complex sentences which basically has a single and a two-sentence conjunctions.
2. Rules Surface Expression can be used to describe the word before and after the conjunctive. So that the compound sentence can be simplified by appropriate because it does not change the meaning and information after simplifying complex sentences.
3. Sentence of 60 samples were available, the percentage of complex sentences simplification results in Indonesian using Rule-Based Reasoning on software as much as 93.3% of the 60 samples in which the existing manjemuk sentences, compound sentences there are four samples that can not be simplified appropriate. This
4. is because an error occurred while defining the token word and sample sentences compound does not have a compound sentence patterns that have been defined.
5. Results simplification of complex sentences are split into two single sentences and the conjunctive word is determined by the class defined. Just a word class of each word in a sentence compound sentence is used to simplify the process of using Rule-Based Reasoning. Therefore, the software can simplify complex sentences are not appropriate when an error in the definition of the word class by NLP_ITB package.

REFERENCES

[1] A. Siddharthan, "An Architecture for A Text Simplification System," in *Language Engineering Conference, 2002. Proceedings*, 2002, pp. 64-71.

[2] N. Yusliani, "Sistem Tanya-Jawab Bahasa Indonesia untuk 'Non-Factoid Question'," Master, Program Studi Informatika, Institut Teknologi Bandung, Bandung, 2010.

[3] V. Seretan, "Acquisition of Syntactic Simplification Rules for French," 2012.

[4] G. G. Chowdhury, "Natural Language Processing," *Annual Review of Information Science and Technology (ARIST),* vol. 37, 2003.

[5] R. A. Sukamto, "Penguraian Bahasa Indonesia dengan Menggunakan Penguraian Collins," Magister, Program Magister Informatika, Institut Teknologi Bandung, Bandung, 2009.

[6] C. Triawati, "Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi," Informatika, ITTELKOM, Bandung, 2009.

[7] H. M. L. S. Jani, Peck, "Applying Machine Learning Using Case-Based Reasoning (CBR) and Rule-Based Reasoning (RBR) Approaches to Object-Oriented Application Framework Documentation," in *Information Technology and Applications, 2005. ICITA 2005. Third International Conference on*, 2005, pp. 52-57 vol.1.

[8] A. Chaer, *Sintaksis Bahasa Indonesia: Pendekatan Proses*: Rineka Cipta, 2009.

[9] B. S. R. Chandrasekar "Automatic Induction of Rules for Text Simplification," *Institute for Research in Cognitive Science,* 1996.

[10] P. M. Nugues, *An Introduction to Language Processing with Perl and Prolog*. Germany: Springer-Verlag Berlin Heidelberg, 2006.

[11] W. Duch, "Rule-Based Methods," Department of Informatics, Nicolaus Copernicus University, Poland, 2010.

[12] S. D. HasanAlwi, Hans Lapoliwa, Anton M. Moelino, "Tata Bahasa Baku Bahasa Indonesia," vol. EdisiKetiga, ed. Jakarta: PusatBahasadanBalaiPustaka, 2003, p. 475.

[13] A. O. Hatem, N. Shaker, "Morphological Analysis for Rule-Based Machine Translation," in *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, 2011, pp. 260-263.

[14] R. Ismoyo, Nasarius Sudaryono, *Bahasa Indonesia untuk Sekolah Dasar/MI Kelas 6*. Jakarta: Pusat

APPENDIX A

Surface Expression Rules in Simplification Complex Sentences

	Conjunction	Before Conjuntion Word	After Conjunction Word	Information
1.	Dan	Objek (noun)	Predikat (verb)	Conjunction on middle of complex sentences
2.	Dan	Subjek (noun)	Subjek (noun)	Conjunction on middle of complex sentences
3.	Dan	Objek (noun)	Objek (noun)	Conjunction on middle of complex sentences
4.	Dan	Predikat (verb)	Predikat (verb)	Conjunction on middle of complex sentences
5.	Tetapi	Predikat (kata kerja)	Predikat (verb)	Conjunction on middle of complex sentences
6.	Tetapi	Objek (noun)	Pelengkap (noun)	Conjunction on middle of complex sentences
7.	Tetapi	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
8.	Tetapi	Objek (noun)	Predikat (verb)	Conjunction on middle of complex sentences
9.	Tetapi	Predikat (verb)	Subjek (noun)	Conjunction on middle of complex sentences
10.	Tetapi	Keterangan (noun)	Keterangan (noun)	Conjunction on middle of complex sentences
11.	Jika	-	Subjek (noun)	Kata penghubung di awal kalimat
12.	Melainkan	Predikat (kata kerja)	Predikat (verb)	Conjunction on middle of complex sentences
13.	Melainkan	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
14.	Melainkan	Objek (noun)	Objek (noun)	Conjunction on middle of complex sentences
15.	Bahkan	Predikat (verb)	Predikat (verb)	Conjunction on middle of complex sentences
16.	Atau	Predikat (verb)	Predikat (verb)	Conjunction on middle of complex sentences
17.	Atau	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
18.	Atau	Objek (noun)	Objek (noun)	Conjunction on middle of complex sentences
19.	Biarpun	-	Subjek (noun)	Kata penghubung di awal kalimat
20.	Jangankan	-	Predikat (verb)	Kata penghubung di awal kalimat
21.	Sedangkan	Objek (noun)	Subjek (noun)	Conjunction on middle of complex sentences
22.	Sedangkan	Predikat (verb)	Subjek (noun)	Conjunction on middle of complex sentences
23.	Karena	Predikat (verb)	Predikat (verb)	Conjunction on middle of complex sentences
24.	Karena	Predikat (verb)	Subjek (noun)	Conjunction on middle of complex sentences
25.	Karena	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
26.	Karena	-	Predikat (verb)	Conjunction in Front of Sentences
27.	Daripada	-	Subjek (noun)	Conjunction in Front of Sentences
28.	Maka	Predikat (verb)	Subjek (noun)	Conjunction on middle of complex sentences
29.	Sehingga	Predikat (verb)	Subjek (noun)	Conjunction on middle of complex sentences
30.	Sehingga	Predikat (verb)	Predikat (verb)	Conjunction on middle of complex sentences
31.	Sehingga	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
32.	Sehingga	Objek (noun)	Predikat (verb)	Conjunction on middle of complex sentences
33.	Saat	-	Subjek (noun)	Conjunction in Front of Sentences
34.	Kemudian	Keterangan (noun)	Predikat (verb)	Conjunction on middle of complex sentences
35.	Kemudian	Objek (noun)	Predikat (verb)	Conjunction on middle of complex sentences
36.	Meskipun	-	Subjek (noun)	Conjunction in Front of Sentences
37.	Meskipun	-	Predikat (verb)	Conjunction in Front of Sentences
38.	Lalu	Objek (noun)	Predikat (verb)	Conjunction on middle of complex sentences
39.	Ketika	-	Subjek (noun)	Conjunction in Front of Sentences
40.	Walaupun	-	Subjek (noun)	Conjunction in Front of Sentences
41.	Walaupun	-	Objek (noun)	Conjunction in Front of Sentences
42.	Agar	Objek (noun)	Predikat (verb)	Conjunction in Front of Sentences