Paraphrasing Method Based on Contextual Synonym Substitution

Ari Moesriami Barmawi[1,*] & Ali Muhammad[2]

[1]Graduate School of Informatics, School of Computing, Telkom University, Kawasan
Pendidikan Telkom, Sukapura, Kec. Dayeuhkolot, Bandung, 40257, Indonesia
[2]Department of Information Technology Education, STKIP PGRI Banjarmasin, Jalan
Adam Sultan complex H. Iyus No. 18 Rt 23, Banjarmasin, Sungai Jingah, Kec.
Banjarmasin Utara, Kota Banjarmasin, Kalimantan Selatan 70121, Indonesia
*E-mail: mbarmawi@melsa.net.id

Abstract. Generating paraphrases is an important component of natural language processing and generation. There are several applications that use paraphrasing, for example linguistic steganography, recommender systems, machine translation, etc. One method for paraphrasing sentences is by using synonym substitution, such as the NGM-based paraphrasing method proposed by Gadag, *et al.* The weakness of this method is that ambiguous meanings frequently occur because the paraphrasing process is based solely on n-gram. This negatively affects the naturalness of the paraphrased sentences. For overcoming this problem, a contextual synonym substitution method is proposed, which aims to increase the naturalness of the paraphrased sentences. Using the proposed method, the paraphrasing process is not only based on n-gram but also on the context of the sentence such that the naturalness is increased. Based on the experimental result, the sentences generated using the proposed method had higher naturalness than the sentences generated using the original method.

Keyword: *context; language; paraphrasing; synonym; substitution.*

1 Introduction

Generating paraphrases is an important component of natural language processing and generation. There are several applications that use paraphrasing, for example linguistic-based steganography, recommender systems, machine translation, etc. One method for paraphrasing sentences is synonym substitution. There are several methods of synonym substitution, such as DIRT [1], bilingual pivoting [2-4], PPDB [5] and NGM-based methods [6]. As the latest method for paraphrasing based on synonym substitution, Gadag's method [6] has problems with the naturalness of the paraphrased sentences. For increasing the naturalness of the paraphrased sentences, a paraphrasing method based on contextual synonym substitution is proposed.

Received March 31st, 2019, Revised December 29th, 2019, Accepted for publication December 31st, 2019.

The paraphrasing process in the proposed method is not based solely on n-gram, but also on the context of the sentence such that the naturalness of the paraphrased sentence is increased. For observing the naturalness of the paraphrased sentence using the original and the proposed method, two experimental scenarios were planned. The first evaluation was metric and used Meteor, while the second one used human judgment. Based on the experimental result, the naturalness of the paraphrased sentences using the proposed method was higher (better) than Gadag's one.

2 N-gram Based Paraphrase Generator

Gadag, *et al.* [8] developed a paraphrase generator method that uses n-gram induction for finding paraphrase word candidates. In the case of n = 3 (3-grams), this method compares the 3-grams in a sentence. Suppose we have N number of 3-grams in the corpus (called the 3-gram set of the corpus), then each sequence of each 3-gram y_i is stored as sequence c_i. All 3-grams are filtered based on the number of common words in the corpus and the paraphrase candidate. In this case at least two common words are required. Suppose one of the filtered 3-grams has a sequence *a*, then the pair of sequences (a, y_i) and (a, y_j) are compared. Finally, if y_i and y_j have at least two common words, then *yi* and y_j are included in a subset of the 3-gram set. An overview of Gadag's method is shown in Figure 1.

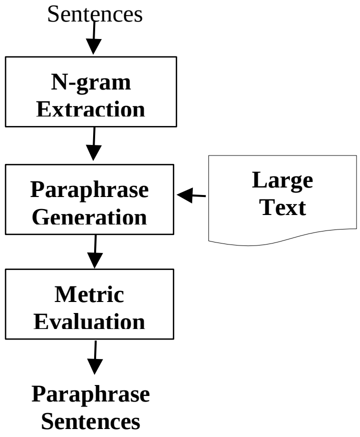


Figure 1 n-gram based paraphrase generator process [6].

Gadag's method [6] uses R-precision metric to evaluate the result of the paraphrased sentence. The precision metric is formulated as follows:

$$R_{precision} = \frac{Number\ of\ overlapping\ words}{Length\ of\ reference\ paraphrase} \quad (1)$$

where the number of overlapping words is the number of words that appear in both the paraphrased candidate and the reference paraphrase, and the length of the reference paraphrase is the total number of words in the reference paraphrase. Based on the evaluation method it was shown that the precision of this method is about 46.3%, which is low.

For example:

Original sentence: *Penjual tahu yang dibutuhkan pembeli*

Trigram:

penjual tahu yang
tahu yang dibutuhkan
yang dibutuhkan pembeli

The corpus is as follows:

1. *Penjual anggur yang sakit. Trigram: penjual anggur yang; anggur yang sakit*
2. *Penjual mengerti yang diperlukan pembeli. Trigram: penjual mengerti yang; mengerti yang diperlukan; yang diperlukan pembeli*
3. *Amin tahu kalau dia diperlukan. Trigram: Amin tahu kalau; tahu kalau dia; kalau dia diperlukan*
4. *Hari paham yang diperlukan konsumennya. Trigram: Hari paham yang; paham yang diperlukan, yang diperlukan konsumennya.*

Since the first trigram of the second message in the corpus and the first trigram of the original sentence have two common words, and the second word of the second message in the corpus '*mengerti*' or '*paham*' is a synonym of the second word of the original message, which is '*tahu*'.

Thus, the paraphrased sentence is

Penjual tahu yang diperlukan pembeli ➔ *Penjual mengerti yang diperlukan pembeli* or

Penjual tahu yang diperlukan pembeli ➔ *Penjual paham yang diperlukan pembeli*

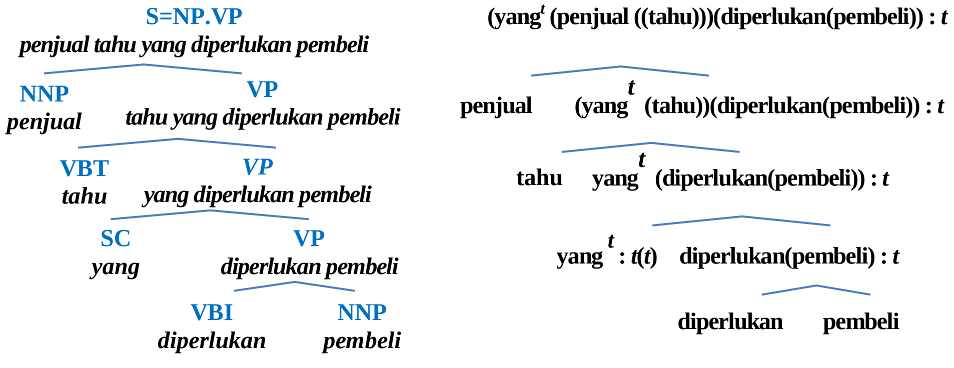
After parsing, the original sentence structure is as follows:

***Penjual/*NNP *tahu/*NN *yang/*SC *diperlukan/*VBI *pembeli/*NNP**

Meanwhile, suppose the paraphrased sentence structure is as follows:

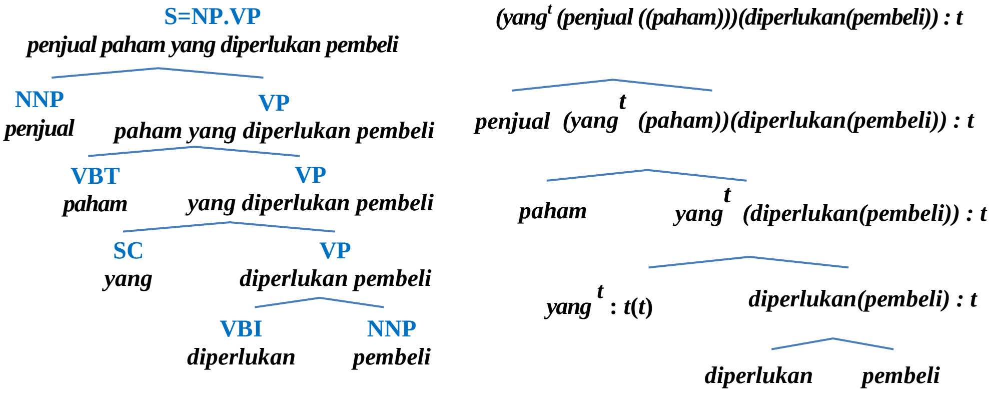
***Penjual*/NNP *mengerti*/VBI *yang*/SC *diperlukan*/VBI *pembeli*/NNP**

Since the paraphrased sentence and the original one have a different structure, the meanings of these two sentences are not the same. Based on the example, the paraphrased sentence using Gadag's method [6] is out of context. The problem occurred because the n-gram method chooses the paraphrase candidates without considering the context of the sentence. The syntactic structure and semantic interpretation of '*penjual tahu yang diperlukan pembeli*' and '*penjual paham yang diperlukan pembeli*' are shown in Figures 2 and 3 respectively.


S=NP.VP
penjual tahu yang diperlukan pembeli
NNP
penjual
VP
tahu yang diperlukan pembeli
VBT
tahu
VP
yang diperlukan pembeli
SC
yang
VP
diperlukan pembeli
VBI
diperlukan
NNP
pembeli
(yang^{t} (penjual ((tahu)))(diperlukan(pembeli)) : t
penjual
(yang^{t} (tahu))(diperlukan(pembeli)) : t
tahu
yang^{t} (diperlukan(pembeli)) : t
yang^{t} : t(t)
diperlukan(pembeli) : t
diperlukan
pembeli

(a) (b)

Figure 2 (a) Syntactic structure and (b) semantic interpretation of '*penjual tahu yang diperlukan pembeli*' using Gadag's method.


S=NP.VP
penjual paham yang diperlukan pembeli
NNP
penjual
VP
paham yang diperlukan pembeli
VBT
paham
VP
yang diperlukan pembeli
SC
yang
VP
diperlukan pembeli
VBI
diperlukan
NNP
pembeli
(yang^{t} (penjual ((paham)))(diperlukan(pembeli)) : t
penjual (yang^{t} (paham))(diperlukan(pembeli)) : t
paham
yang^{t} (diperlukan(pembeli)) : t
yang^{t} : t(t)
diperlukan(pembeli) : t
diperlukan
pembeli

(a) (b)

Figure 3 (a) Syntactic structure and (b) semantic interpretation of '*penjual paham yang diperlukan pembeli*' (paraphrased sentence) using Gadag's method.

3 Indonesian Grammar

The basic sequence for Indonesian sentences is *subject* followed by *predicate*, and *object*, *complement*, or *modifier*. The tag set used in Indonesian language is shown in Table 1, and POS forming phrases are shown in Table 2.

Table 1 Tag Set in Indonesian language [7].

POS	POS Decription	Example
OP	Open Parenthesis	({[
CP	Close Parenthesis	)}]
GM	Slash	/
;	Semicolon	;
:	Colon	:
	Quotation	*
.	Sentence Terminator	.!?
,	Comma	,
-	Cash	-
...	Ellipsis	...
JJ	Adjective	*Cantik, cepat*
RB	Adverb	*Sementara, nanti*
NN	Common Noun	*Sepeda*
NNP	Proper Noun	*Jakarta, Bandung*
NNG	Genitive Noun	*Kursinya*
VBI	Intransitive Verb	*Datang*
VBT	Transitive Verb	*Menjual*
IN	Preposition	*Ke, di, pada*
MD	Modal	*Bisa*
CC	Co or-Conjunction	*Atau, tetapi, dan*
SC	Sub or-Conjunction	*Jika, ketika*
DT	Determiner	*Ini, itu, para*
UH	Interjection	*Wah, aduh, oi*
CDO	Ordinal Numerals	*Ketiga, keempat*
CDC	Collective Numerals	*Berempat*
CDP	Primary Numerals	*Tiga, empat*
CDI	Irregular Numerals	*Beberapa*
PRP	Personal Pronouns	*Saya, kamu*
WP	WH-Pronouns	*Apa, siapa*
PRN	Number Pronouns	*Kedua-duanya*
PRL	Locative Pronouns	*Sini, situ, sana*
Neg	Negation	*Tidak, bukan*
SYM	Symbols	@#$%^&
RP	Particles	*Pun, kah*
FW	Foreign Words	Foreign, word

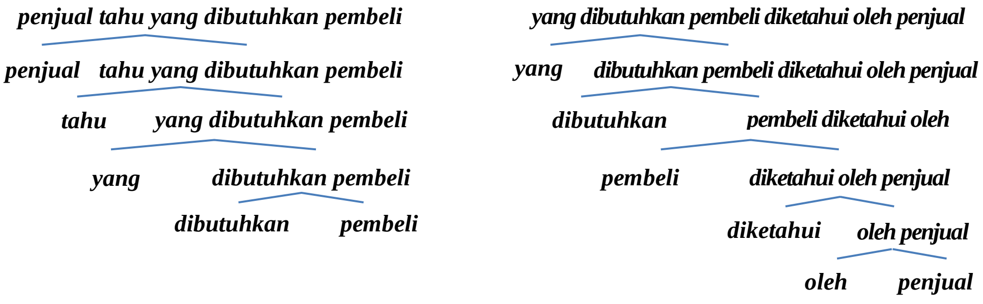
Table 2 POS forming phrases [7].

Types of Phrases	POS Tag
Questioning Phrase (TP)	WP
Numeric Phrases (BP)	CDO, CDP, CDI, CDC
ConnectionPhrases (KP)	SC, CC, NEG, IN, MD, RB
Noun Phrases (NP)	NN, PRN, PRP, PRL, NNG, NNP, FW, RP, UH, JJ
Verb Phrases (VP)	VBI, VBT

Similar to English, in Indonesian language, verbs are classified into transitive and intransitive verbs. Transitive verbs are used in active sentences, while intransitive verbs are used in passive sentences. Examples of transitive and intransitive uses of words are shown in Table 3 and the semantic transformation is shown in Figure 4. It is shown in Table 1 that in the sentence '*Penjual tahu yang diperlukan pembeli*', '*penjual*' is tagged as a personal pronoun (PP), '*tahu*' is tagged as a verb (VBT), '*yang*' is tagged as a conjunction (SC), '*diperlukan*' is tagged as a verb (VBI), and '*pembeli*' is tagged as a noun (NN). Based on Figure 4, it is shown that the syntactical transformation of the active sentence '*penjual tahu yang diperlukan pembeli*' is '*yang diperlukan pembeli diketahui penjual*'.

Table 3 Examples of Indonesian transitive and intransitive words.

Transitive	Intransitive
Dia akan menyetir mobil	*Dia ditinggalkan oleh temannya*
PRP RB VBT NN	PRP VBI IN NNG
Dia sudah menyetir mobil	*Pesawat sedang mendarat*
PRP RB VBT NN	NN RB VBI
Saya menyetir mobil	*Anak itu ketiduran*
PRP VBT NN	NN DT VBI



(a) (b)

Figure 4 Syntactical transformation from (a) active to (b) passive.

4 Contextual Synonym Substitution

In this case, a synonym word list is used as the input of the contextual substitution method. There are two processes in the proposed method, word filtering and contextual synonym substitution. An overview of the proposed method is shown in Figure 5.

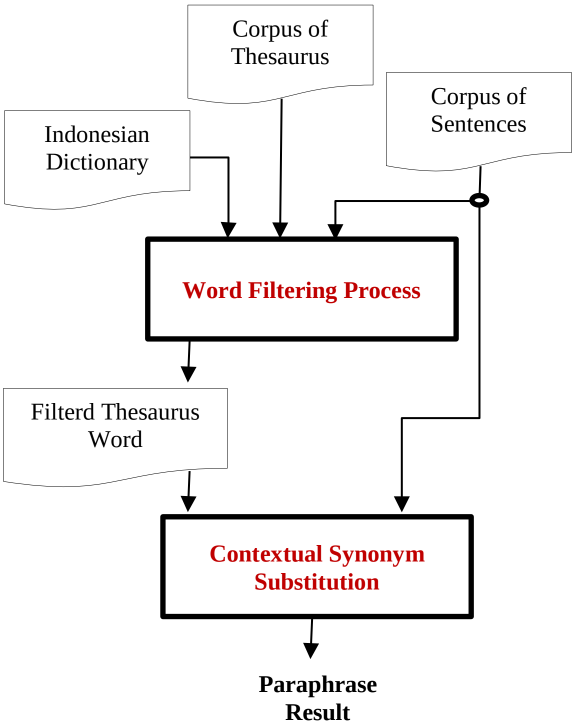


Figure 5 Proposed method.

4.1 Word Filtering Process

Word filtering is a process for filtering words in the Indonesian dictionary and thesaurus based on the synonyms of words used in the corpus. This process is necessary since not all words in Indonesian dictionary and thesaurus are used in the corpus.

Three processes are conducted for word filtering: POS tagging, word and tag filtering, and thesaurus word filtering. An overview of the word filtering process is shown in Figure 6.

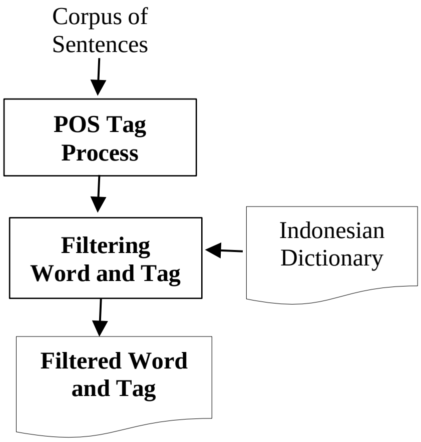


Figure 6 Word filtering process.

4.1.1 POS Tagging Process

POS tagging is used to assign a word class to each word based on its syntactic function (POS: part of speech), such as noun, verb, adverb, adjective, etc. For POS tagging, hidden Markov model [8] is used. POS tagging involves tokenization of the words in the Indonesian dictionary. This method uses probabilistic and state transition diagram (rule based) methods to determine the right tag for a word.

The words are separated into two classes: closed-class words and open-class words. Open-class words are words such as nouns, verbs and adjectives, while closed-class word are words such as pronouns and conjunctions. The POS tagging process is started by processing the closed-class words, followed by processing the open-class words. When ambiguity occurs, a predefined rule was used to find the right word tag.

Example of tagging process for the Indonesian language:

saya anak nakal sekali **→ *saya*/PRP *anak*/NN *nakal*/JJ *sekali*/RB**

Example of tagging process for the English language:

Their job is to make it so compellingly obvious that one day everyone sees it → their/PRP job/NN is/VBZ to/TO make/VB it/PRP so/RB compellingly/RB obvious/JJ that/IN one/CD day/NN everyone/NN sees/VBZ it/PRP

4.1.2 Filtering Word and Tag

After conducting POS tagging, the word and its tag are filtered based on the word and its tag in Indonesian dictionary. For filtering existing words and tags, this method compares a word in the corpus and its tag with the same word and tag in Indonesian dictionary. Suppose we have the word ‘*nakal*’ in the corpus, tagged as JJ, then the word ‘*nakal*’ and the tag JJ have to be found in Indonesian dictionary. When the word ‘*nakal*’ and the tag are found in the dictionary, the word ‘*nakal*’ and its tag JJ are included into a specific list of words and tags. Otherwise, the word ‘*nakal*’ is excluded from the word list. The same process is applied to English sentences. In this case, the word ‘make’, tagged as VB is chosen to be filtered. Then, the same process as in Indonesian language is applied to the word ‘make’.

4.2 Context-based Substitution Process

Context-based substitution is conducted to substitute words in the corpus with its synonym. The inputs of this process is a synonym word list and a corpus. There are three sub-processes of the context-based substitution process: POS tagging, substitution, and structural evaluation. An overview of the context-based substitution process is shown in Figure 7.

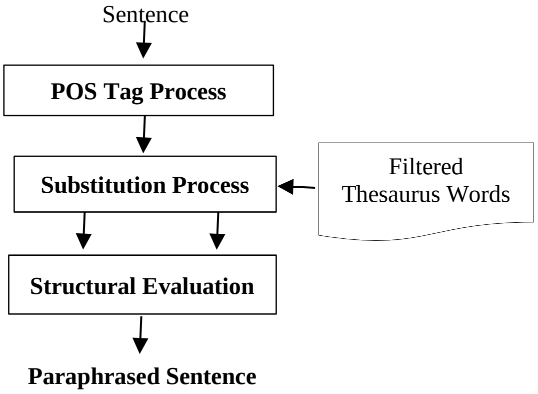


Figure 7 Contextual synonym substitution.

4.2.1 Substitution Process

After conducting POS tagging on the corpus (as discussed in 4.1.1), the words used in the corpus are substituted based on the filtered thesaurus word list. This method generates possible substitution words that can be used to substitute words used in sentences in the corpus.

The synonym substitution method used in this process is based on interchangeable words, i.e. words that are interchangeable between one another, even for different meanings [9]. To find interchangeable words, a synonym set is generated based on the word occurrence in the corpus by calculating the occurrence probability of a word [10]. The occurrence frequency (f_n) of the word '*diperlukan*' in the corpus (as in the example in Section 2) is shown in Table 4. In this case, for 2-grams, 3-grams, and 4-grams, f_n is equal to 5. Then, the count function (*Count*(w)), (where w is the word '*diperlukan*') is calculated using Eq. (2).

$$Count(w) = \sum_2^m \log(f_n) \tag{2}$$

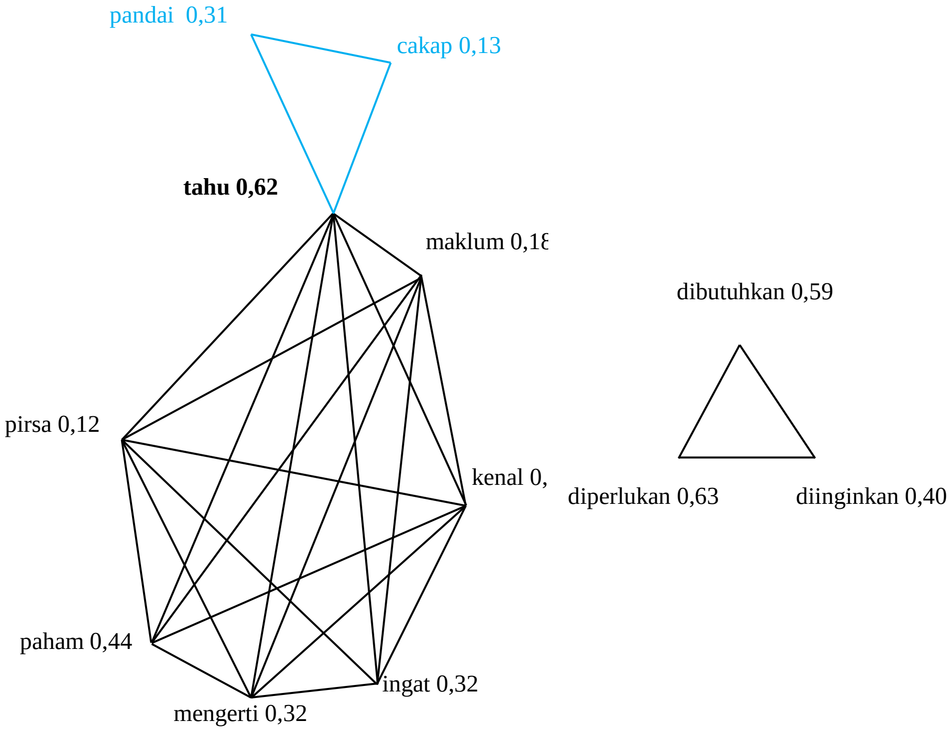
Table 4 N-gram of Corpus (example in Section 2) and its frequency .

n-gram	Frequency	f_n
yang ***diperlukan***	2	5
dia ***diperlukan***	1	
diperlukan *pembeli*	1	
diperlukan *konsumennya*	1	
yang ***diperlukan*** *pembeli*	1	5
yang ***diperlukan*** *konsumennya*	1	
mengerti yang ***diperlukan***	1	
kalau dia ***diperlukan***	1	
paham yang ***diperlukan***	1	
Penjual mengerti yang ***diperlukan***	1	5
tahu kalau dia ***diperlukan***	1	
mengerti yang ***diperlukan*** *pembeli*	1	
Hari paham yang ***diperlukan***	1	
paham yang ***diperlukan*** *konsumennya*	1	

Furthermore, the maximum value of *Count(w)* (called max_{count}) is obtained, continued by calculating the proportion between the count of '*diperlukan*' and max_{count}. This proportion is called $score_{\mathrm{NGM}}$(*diperlukan*), as shown in Eq. (3).

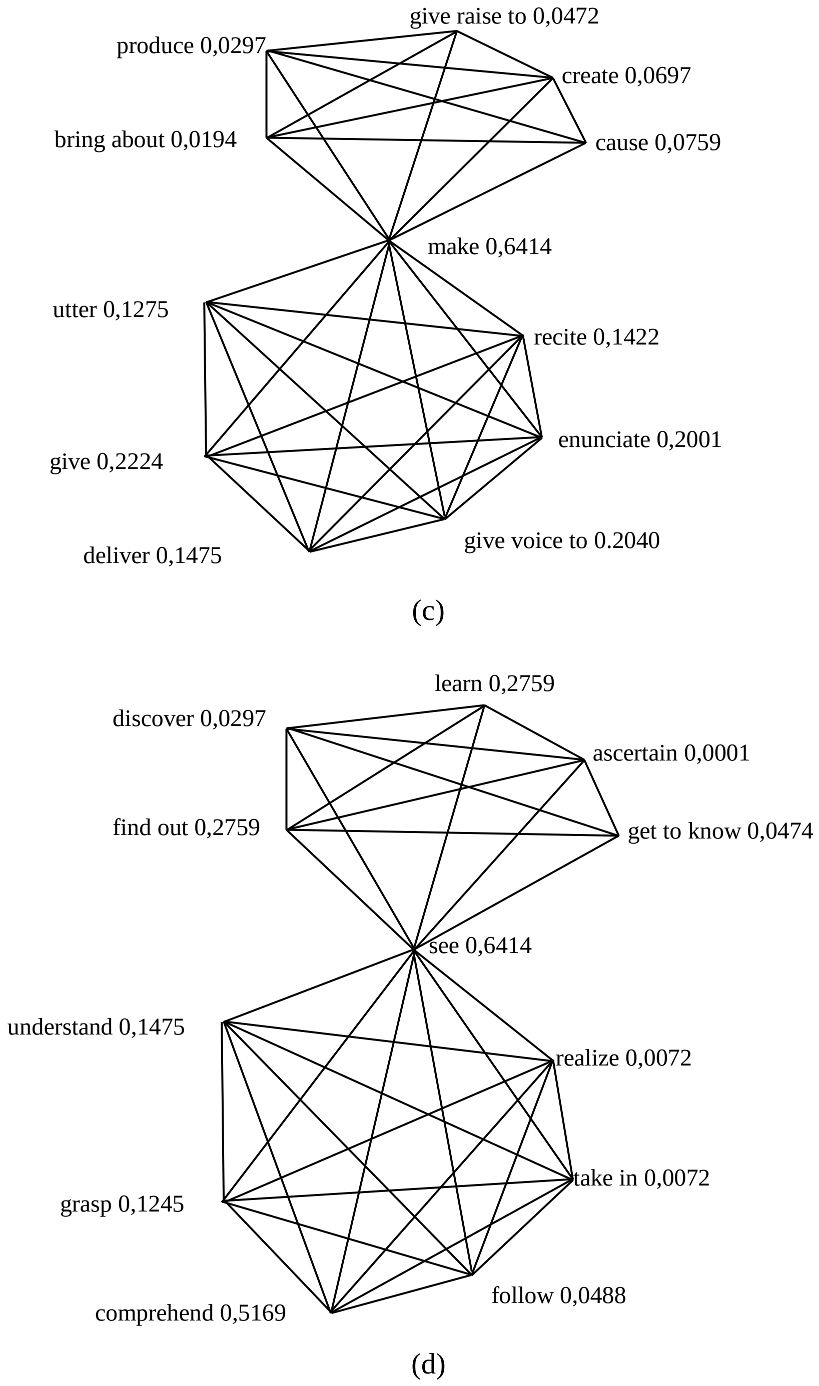
$$score_{NGM}(w) = \frac{count(w)}{max_{count}} \tag{3}$$

score$_{NGM}$ is used in the synonym generation process. For generating the synonym set (synset), a *score*$_{NGM}$ threshold is determined. Words with *score*$_{NGM}$ greater than or equal to the threshold are included in the graph of the synset, otherwise they are removed from the graph. The same process is applied to the English words 'go' and 'see'. An example of synonym substitution in the Indonesian language is shown in Figure 8(a) and (b). Based on Figure 8(a) and (b), the word '*tahu*' has two synsets. The first synset is {*cakap, pandai,tahu*} and the second one is {*pirsa, paham, mengerti, ingat, kenal, maklum*}. The word '*diperlukan*' has one synset {*dibutuhkan, diinginkan, diperlukan*}. The value following the word is the value of *score*$_{NGM}$. Suppose the threshold applied in the synset of '*tahu*' is 0.3 and 0.4 in the synset of '*diperlukan*', then all words in the synset of '*tahu*' whose *score*$_{NGM}$ $\leq$ 0.3 are removed and all words in the synset of '*diperlukan*' whose *score*$_{NGM}$ $\leq$ 0.4 are removed.



(a) (b)

Figure 8 Synonym graph of words (a) '*tahu*', (b) '*diperlukan*'.



(c)

(d)

Figure 8 ***Continued.*** Synonym graph of words, (c) ‘make’, (d)’see’.

The synonym graph after removing the words that did not fulfill the requirements, is shown in Figures 9(a) and (b). The higher *score*$_{NGM}$, the larger the probability of generating a natural sentence. In the synset of '*tahu*', the word '*pandai*' had score$_{\text{NGM}}$ = 0.31, which is greater than 0.3, but this word was removed because it is an adjective and not a verb, so that it is not interchangeable with '*tahu*'. In other words, this word is not in the same context. Suppose the word '*diperlukan*' is used to substitute the word '*dibutuhkan*' and the word '*paham*' is used to substitute the word '*tahu*'. Thus the paraphrased sentence becomes '*penjual paham apa yang dibutuhkan pembeli*'.

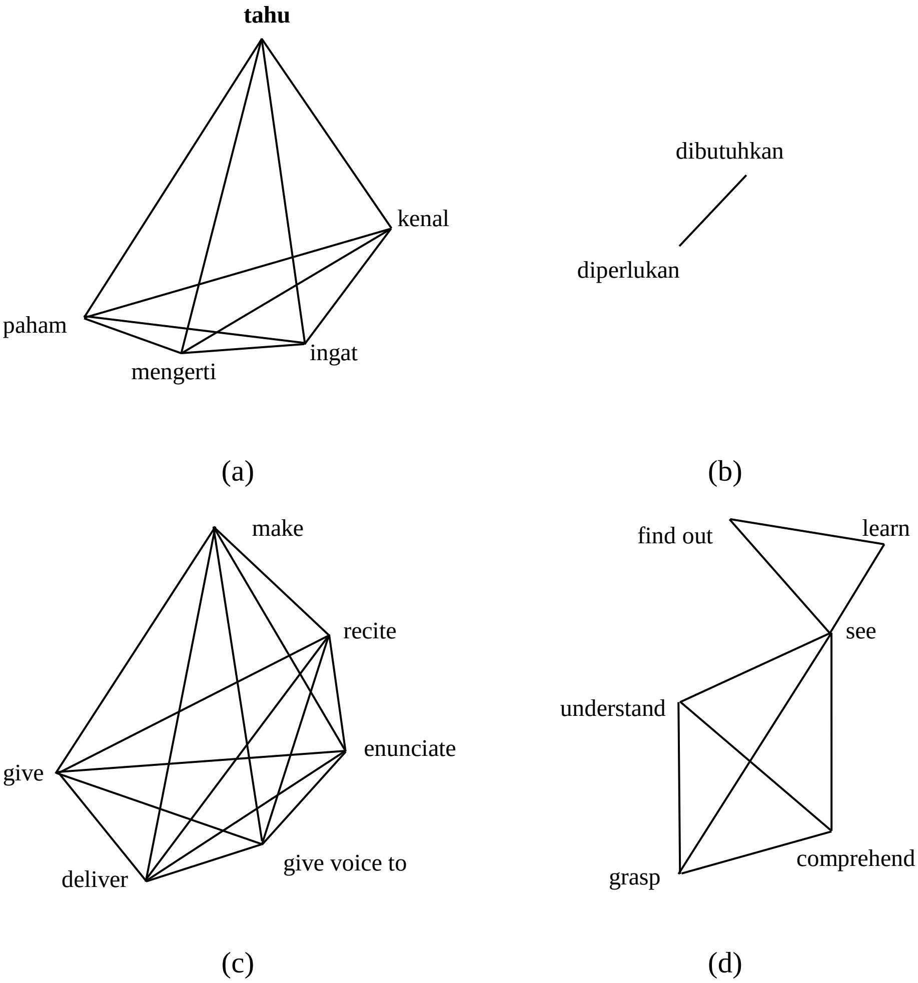


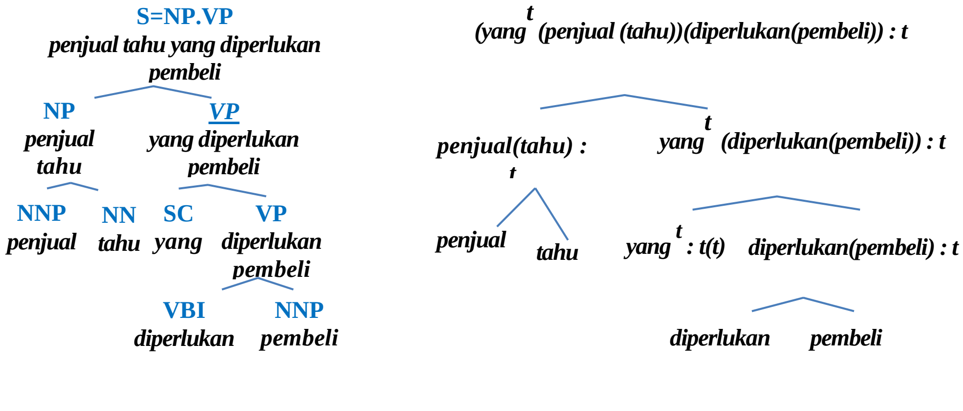
Figure 9 Synonym graphs of words (a) '*tahu*', (b) '*diperlukan*', (c) 'make', (d)'see', after word removal.

In the case of English (as shown in Figures 8(c) and (d)), the words ‘make’ and ‘see’ each have a synset. These synsets are generated after removing words that are not a verb. The thresholds of the ‘make’ and ‘see’ synsets were 0.1300, and 0.1200 respectively. The synonym graph after the threshold-based words have been removed is shown in Figures 9(c) and (d). Suppose the word ‘deliver’ is used to substitute the word ‘make’ and the word ‘comprehend’ is used to substitute the word ‘see’, then the paraphrased sentence becomes ‘Their job is to deliver it so compellingly obvious that one day everyone comprehends it’. After implementing the word substitution, the paraphrased sentences are grouped into a paraphrased sentence candidate list.

4.2.2 Structural Evaluation

Structural evaluation is a method to filter the sentence candidate list (in the corpus) for context-based paraphrasing based on the sequence tag of the original sentence. In this case, if the sequence tag of the sentences in the sentence candidate list for context-based paraphrasing is the same as the sequence tag of the words in the original sentence, then the sentence in the sentence candidate list can be used as the paraphrased sentence.

Suppose we have the original sentence ‘*penjual tahu yang diperlukan pembeli*’, then the syntactic structure and the semantic interpretation is shown in Figure 10. The POS tagging is *Penjual*/NNP *tahu*/NN *yang*/SC *diperlukan*/VBI *pembeli*/NNP.



(a) (b)

Figure 10 (a) Syntactic structure and (b) semantic interpretation of ‘*Penjual tahu yang diperlukan pembeli*’.

Since the word ‘*tahu*’ was tagged as a verb in the synset while the tag in the original sentence was a noun, then the word ‘*tahu*’ could not be replaced by the words in the synset of ‘*tahu*’. Therefore, the word ‘*tahu*’ could not be substituted. Thus, the word that could be substituted was ‘*diperlukan*’, because the tag of ‘*diperlukan*’ in the synset was the same as the tag of ‘*diperlukan*’ in the original sentence. Finally, the word ‘*dibutuhkan*’ was chosen as the synonym of ‘*diperlukan*’, and the paraphrased sentence was: ‘*penjual tahu yang dibutuhkan pembeli*’. The syntactic structure and semantic interpretation of the paraphrased sentence are shown in Figure 11.

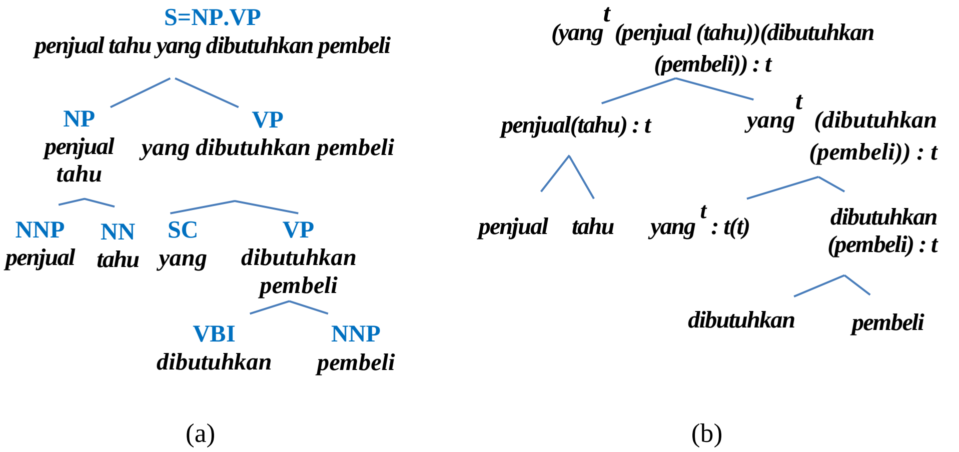
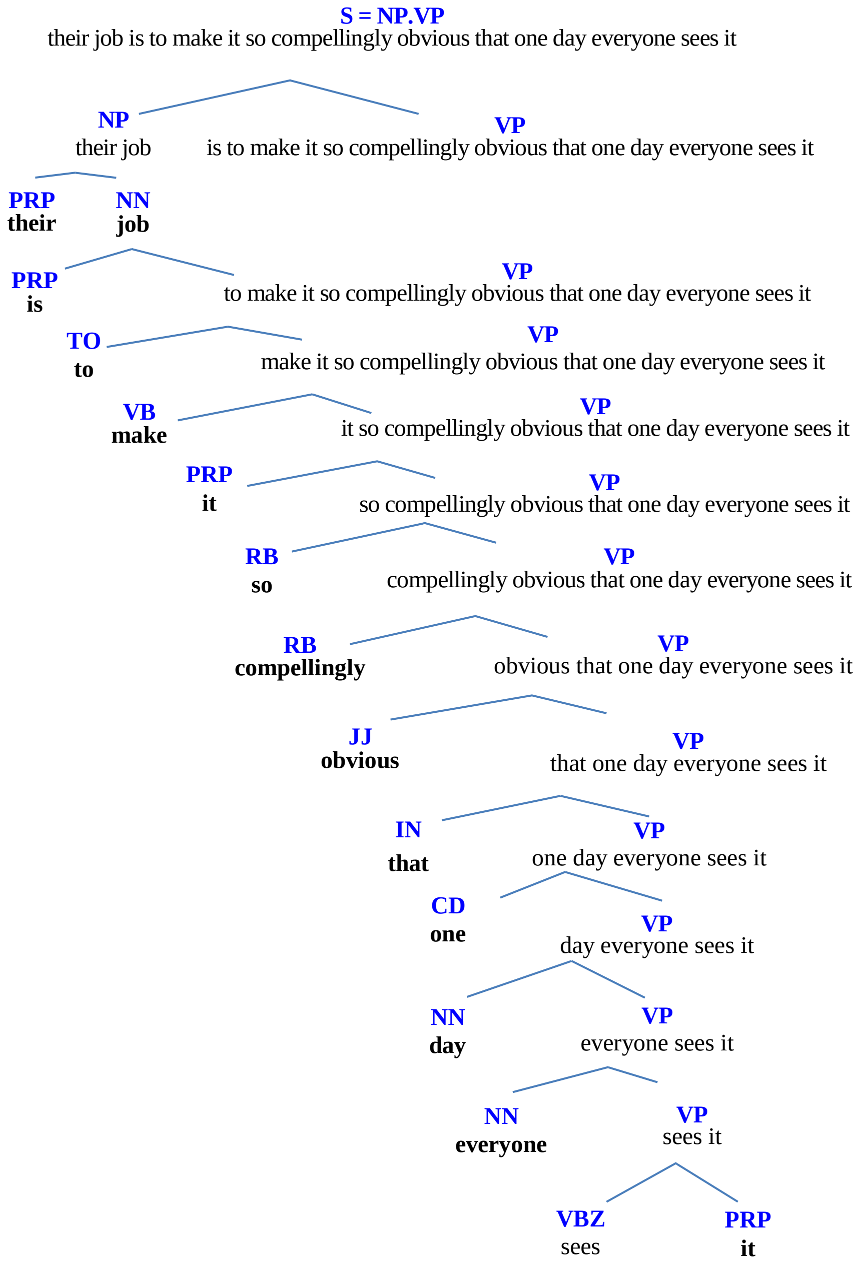


Figure 11 (a) Syntactic structure and (b) semantic interpretation of proposed paraphrased sentence.

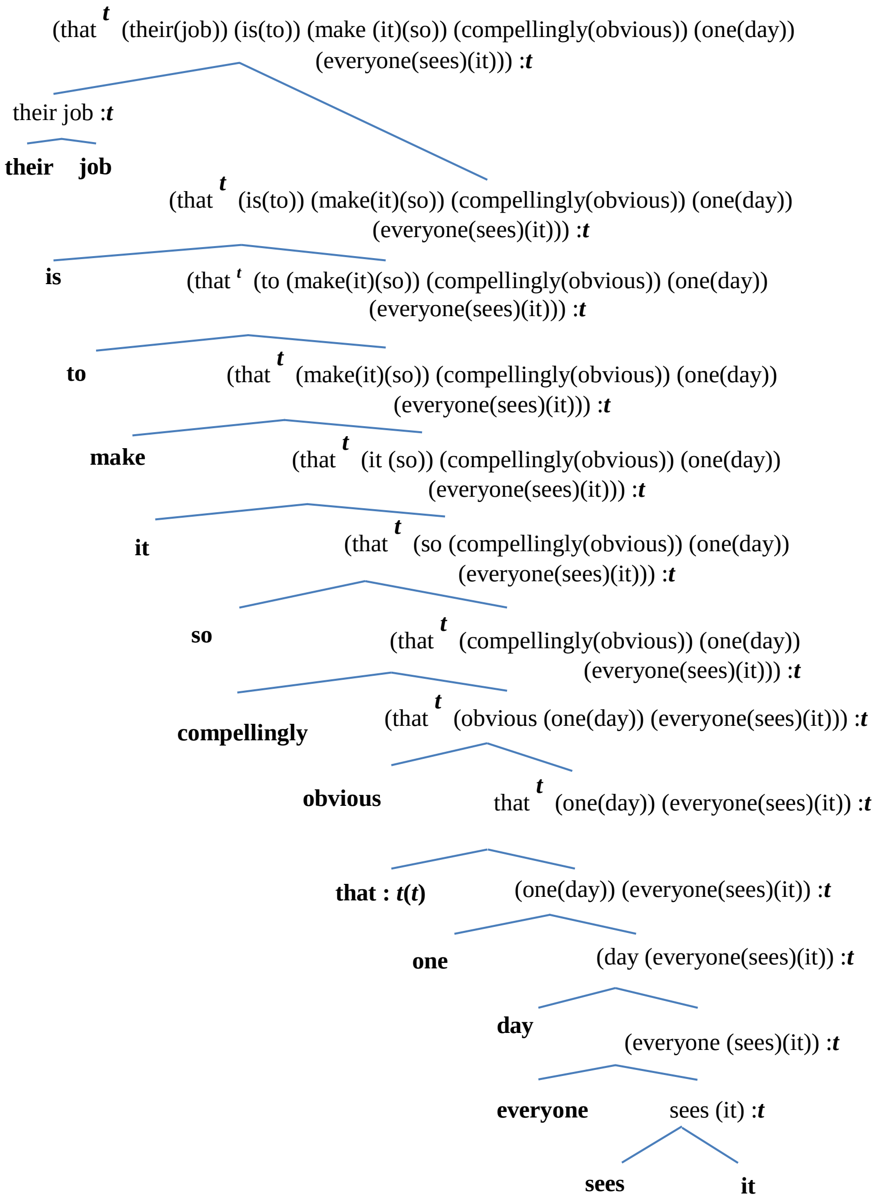
Suppose the sentence ‘Their job is to make it so compellingly obvious that one day everyone sees it’ is used as an example of an English sentence, then the POS tagging is: their/PRP job/NN is/VBZ to/TO make/VB it/PRP so/RB compellingly/RB obvious/JJ that/IN one/CD day/NN everyone/NN sees/VBZ it/PRP (See Figure 12).

Since the words ‘make’ and ‘see’ were tagged as verbs in the synset, which is the same as in the original sentence, these words could be substituted by words in the synsets of ‘make’ and ‘see’. Suppose ‘deliver’ is used to substitute ‘make’ and ‘comprehend’ is used to substitute the word ‘see’, then the paraphrased sentence becomes ‘Their job is to deliver it so compellingly obvious that one day everyone comprehends it’. The syntactic structure and semantic interpretation of the paraphrased sentence are shown in Figure 13.


S = NP.VP
their job is to make it so compellingly obvious that one day everyone sees it
NP
their job
VP
is to make it so compellingly obvious that one day everyone sees it
PRP
their
NN
job
PRP
is
VP
to make it so compellingly obvious that one day everyone sees it
TO
to
VP
make it so compellingly obvious that one day everyone sees it
VB
make
VP
it so compellingly obvious that one day everyone sees it
PRP
it
VP
so compellingly obvious that one day everyone sees it
RB
so
VP
compellingly obvious that one day everyone sees it
RB
compellingly
VP
obvious that one day everyone sees it
JJ
obvious
VP
that one day everyone sees it
IN
that
VP
one day everyone sees it
CD
one
VP
day everyone sees it
NN
day
VP
everyone sees it
NN
everyone
VP
sees it
VBZ
sees
PRP
it

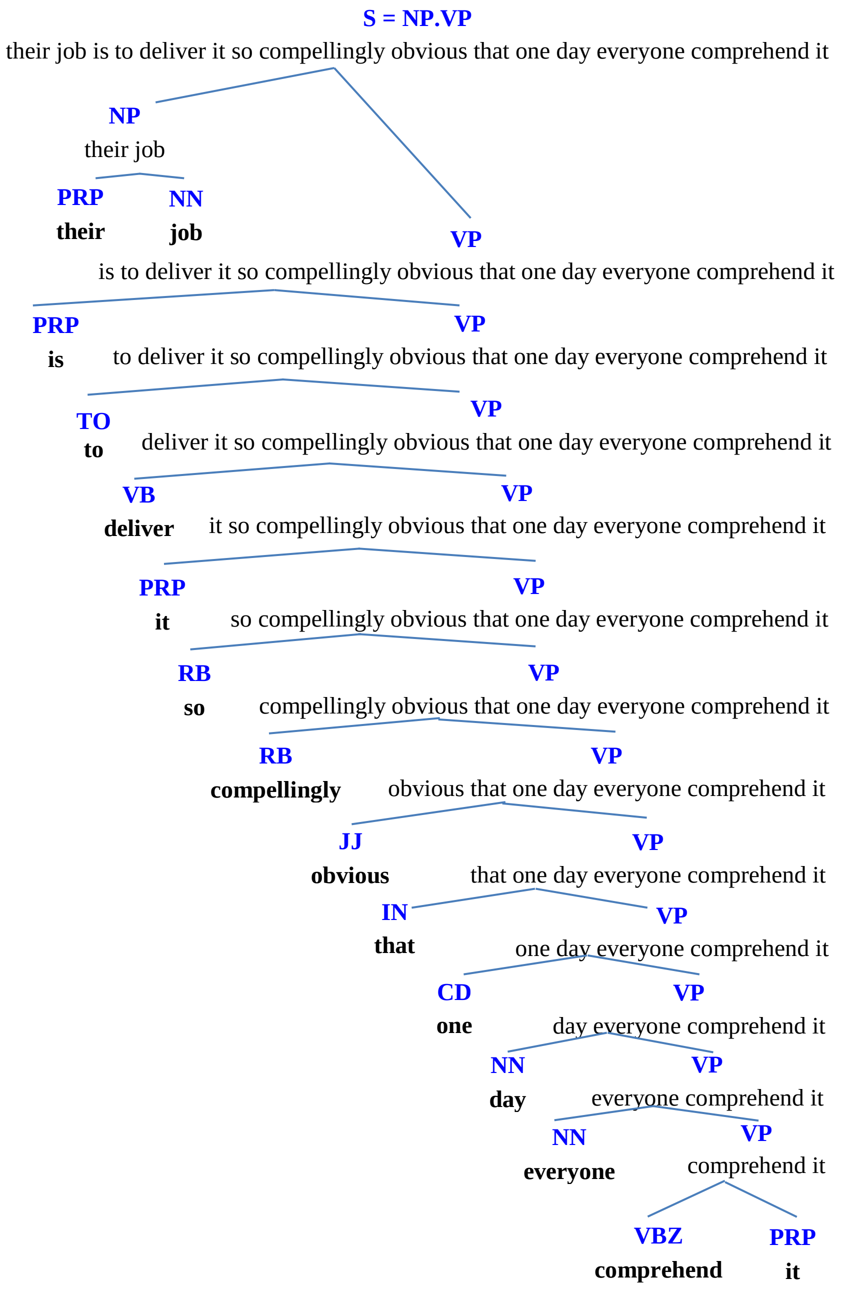
(a)

Figure 12 (a) Syntactic structure of ‘Their job is to make it so compellingly obvious that one day everyone sees it’.


(that t (their(job)) (is(to)) (make (it)(so)) (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
their job :t
their
job
(that t (is(to)) (make(it)(so)) (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
is
(that t (to (make(it)(so)) (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
to
(that t (make(it)(so)) (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
make
(that t (it (so)) (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
it
(that t (so (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
so
(that t (compellingly(obvious)) (one(day)) (everyone(sees)(it))) :t
compellingly
(that t (obvious (one(day)) (everyone(sees)(it))) :t
obvious
that t (one(day)) (everyone(sees)(it)) :t
that : t(t)
(one(day)) (everyone(sees)(it)) :t
one
(day (everyone(sees)(it)) :t
day
(everyone (sees)(it)) :t
everyone
sees (it) :t
sees
it

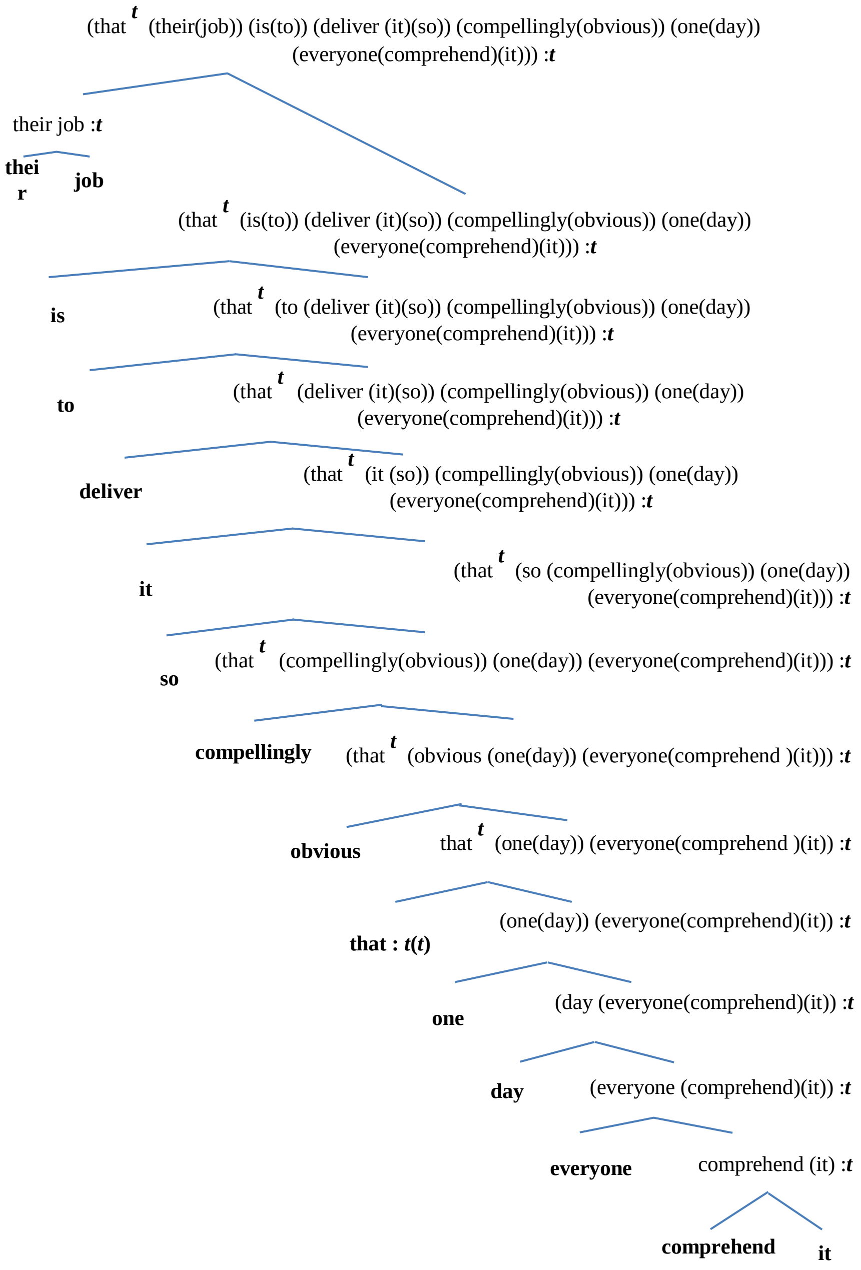
(b)

Figure 12 ***Continued.*** (b) Semantic interpretation of ‘Their job is to make it so compellingly obvious that one day everyone sees it’.


S = NP.VP
their job is to deliver it so compellingly obvious that one day everyone comprehend it
NP
their job
PRP
their
NN
job
VP
is to deliver it so compellingly obvious that one day everyone comprehend it
PRP
is
VP
to deliver it so compellingly obvious that one day everyone comprehend it
TO
to
VP
deliver it so compellingly obvious that one day everyone comprehend it
VB
deliver
VP
it so compellingly obvious that one day everyone comprehend it
PRP
it
VP
so compellingly obvious that one day everyone comprehend it
RB
so
VP
compellingly obvious that one day everyone comprehend it
RB
compellingly
VP
obvious that one day everyone comprehend it
JJ
obvious
VP
that one day everyone comprehend it
IN
that
VP
one day everyone comprehend it
CD
one
VP
day everyone comprehend it
NN
day
VP
everyone comprehend it
NN
everyone
VP
comprehend it
VBZ
comprehend
PRP
it

(a)

Figure 13 (a) Syntactic structure of ‘Their job is to deliver it so compellingly obvious that one day everyone comprehend it’.


(that t (their(job)) (is(to)) (deliver (it)(so)) (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
their job :t
thei r
job
(that t (is(to)) (deliver (it)(so)) (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
is
(that t (to (deliver (it)(so)) (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
to
(that t (deliver (it)(so)) (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
deliver
(that t (it (so)) (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
it
(that t (so (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
so
(that t (compellingly(obvious)) (one(day)) (everyone(comprehend)(it))) :t
compellingly
(that t (obvious (one(day)) (everyone(comprehend)(it))) :t
obvious
that t (one(day)) (everyone(comprehend)(it)) :t
that : t(t)
(one(day)) (everyone(comprehend)(it)) :t
one
(day (everyone(comprehend)(it)) :t
day
(everyone (comprehend)(it)) :t
everyone
comprehend (it) :t
comprehend
it

(b)

Figure 13 ***Continued.*** (b) Semantic interpretation of ‘Their job is to deliver it so compellingly obvious that one day everyone comprehend it’.

5 Experiment and Discussion

This section describes the method and tools for evaluation as well as the experimental result.

5.1 Naturalness Evaluation

As the corpus, this study used the Twitter accounts of Indonesian newspapers, @tempodotco [11], @kompasdotcom [12], @kompascom [13], and @hariankompas. The goal of selecting newspapers twitter accounts was to get formal and natural sentences. To obtain formal and natural sentences, regular expressions such as ‘hastag’, ‘username’, etc. were removed. Two methods were used to evaluate the naturalness of the paraphrased sentences: metric evaluation and human judgment. This study used the twitter accounts from the Indonesian newspapers Pikiran Rakyat and Jawa Pos for testing the naturalness of the sentences generated by the proposed method. The evaluation was done using metric evaluation (Meteor) [14].

5.1.1 Metric Evaluation (Meteor Universal Tool)

For evaluating the performance of the proposed method, evaluation based on n-gram is necessary. Evaluation based on n-gram was done by implementing a penalty. Meteor evaluates a paraphrased sentence by calculating a score based on word-to-word matching between the paraphrased sentence and the reference sentence.

The procedure for evaluating the naturalness of the paraphrased sentence is as follows:

1. Write a list of all possible unigram mappings from the paraphrased word to the reference sentence.
2. Select the largest unigram mapping list, such that one unigram in the paraphrased sentence can be mapped only to one word of the reference sentence.
3. Calculating the precision that represents the accuracy level of the system. The accuracy of the system only considers the number of matched unigrams. The calculation of precision is formulated as follows:

$$P = \frac{\sum_i w_i.(\delta.m_i(h_c)+(1-\delta).m_i(h_f))}{\delta.|h_c|+(1-\delta).|h_f|} \quad (4)$$

where P is precision, Σ_i is the total number of test words, w_i is the observation word, m_i is number of matched paraphrased words, h_c is the content word, and h_f is the function word covered by the matched word in the test sentence, and δ is 10^{-3}.

4. Calculating recall represents the accuracy level of the system to find the word fraction of the paraphrased sentence that appears in the test sentence. The accuracy of the system is only based on the number of matched unigrams. The calculation of recall is formulated as follows:

$$R = \frac{\sum_i w_i.(\delta.m_i(r_c)+(1-\delta).m_i(r_f))}{\delta.|r_c|+(1-\delta).|r_f|} \tag{5}$$

   where R is recall, Σ_i is the total number of test words, w_i is the observation word, m_i is the number of matched test words, r_c is the content word and r_f is the function word covered by the matched word in the paraphrased sentence, and δ is 10^{-3}.
5. Calculating an aggregated score of precision and recall [15], the harmonic mean F1 is:

$$F_1 = \frac{2PR}{P+R} \tag{6}$$

   where P is precision and R is recall.
6. Evaluating the paraphrasing acceptability was done by measuring the similarity of the semantic frames and their role fillers between the reference and the paraphrased sentence, represented by F_{mean}. Meteor is often regarded as a recall-oriented metric; it takes alpha (α) as the relative weight control between precision and recall. The specified alpha value is 0.9 so that the F_{mean} result is concluded as natural language and matched with human judgment perception. F_{mean} is calculated as follows:

$$\boldsymbol{F_{mean}} = \frac{\boldsymbol{P.R}}{\boldsymbol{\alpha.P+(1-\alpha).R}} \tag{7}$$

   where P is precision, R is Recall, and $\alpha = 0.9$ [14].
7. For evaluating the paraphrasing based on n-gram, it is necessary to measure the naturalness of the paraphrased sentence and the correlation between the paraphrased sentence and the reference sentence, which appears to be the same or has the same meaning. The closeness of the sentences' meaning is concluded based on the comparison result between the smallest chunk of the paraphrased sentence and the reference sentence. Chunk is defined as an adjacent and identical sequence between the words in the test sentence and the words in the paraphrased sentence. Suppose we have the test sentence **‘*saya(*i am*) anak(*son*) nakal(*naughty*) sekali(*very*)*’**, and we have the paraphrased sentence **‘*saya*(i am) *anak*(son) *nakal*(naughty) *dan jahat*(bad)’**. Then, we have to find the same sequence words **‘*saya*(i am) *anak*(son) *nakal*(naughty)’** in the paraphrased sentence. The word sequence **‘*saya*(i am) *anak*(son) *nakal*(naughty)’** sliced from the test sentence is called a chunk. The fragmentation penalty is calculated as the number of deductions divided by the number of matched candidate words. Suppose γ sets the maximum penalty and β sets the functional relation

between the fragmentation and the penalty. The fragmentation penalty (*Pen*) is calculated as follows:

$$Pen\ = \gamma\left(\frac{ch}{m}\right)^{\beta} \tag{8}$$

where *ch* is the number of chunks of the test sentence and *m* is the number of matched unigrams, γ is 0.5 and β is 3.0 [16].

8. Finally, calculating the final score to represent the total aggregation value consists of precision, recall, F_{mean}, as shown in Eq. (9).

$$FinalScore = (1 - Pen).F_{mean} \tag{9}$$

5.1.2 Human Judgment Evaluation

This evaluation relies on human perception, which is related to knowledge and experience. There are several methods for human judgment evaluations such, as interviews, questionnaires, and polling.

This study used 56 respondents divided into two groups, namely experts (21 respondents) and non-experts (35 respondents). The experts were journalist and staff from newspapers or mass media, both online and offline, and lecturers of business communication programs. The non-expert respondents were active in accessing, reading and writing newspapers, such as students, lecturers, researchers and ordinary people. Each respondent had to evaluate 25 original sentences compared to 25 paraphrased sentences using Gadag's method and 25 paraphrased sentences using the proposed method. The respondents had to determine the unnaturalness of the sentences and comment on incorrect words. The naturalness is represented by the percentage of naturalness using Eq. (10):

$$Naturalness\ Percentage = \frac{nat_{par}}{nat_{ori}} * 100\% \tag{10}$$

where nat_{par} is the naturalness of the paraphrased sentence and nat_{ori} is the naturalness of the original sentence

5.2 Discussion

This section discusses and analyzes the results from Meteor and from human judgment.

5.2.1 Meteor Universal Tool Result

This study used samples of 100, 500, and 1000 original sentences from news Twitter accounts. Then, paraphrased sentences were generated using Gadag's and the proposed method, respectively, and the naturalness of the results was compared based on the final score resulted from Meteor universal. The result is shown in Figure 14.

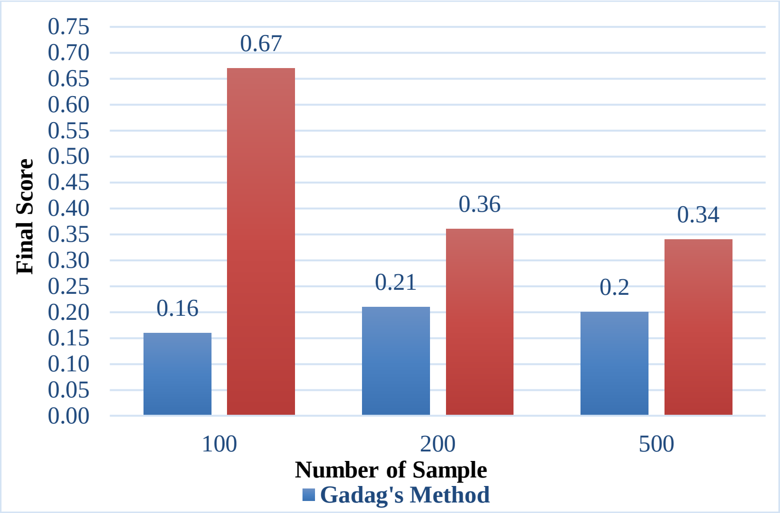


Figure 14 Final score of sentences in the corpus

Based on Figure 14, the naturalness of the paraphrased sentences using Gadag's method was lower than that using the proposed method. This occurred because Gadag's method does not maintain the grammatical structure of the original sentence so that several changes based on n-gram are applied to the original sentence. Meanwhile, the proposed method only changes words based on their synonym word list considering the context. This makes the naturalness of paraphrased sentence using the proposed method higher than that using Gadag's method.

For testing the performance of the proposed method, samples of 20 and 100 sentences were taken from outside of the corpus. For this purpose, original sentences from the Twitter accounts of the newspapers Pikiran Rakyat [17] and Jawa Pos [18] were used. For evaluating the performance of both methods, 20 and 100 paraphrased sentences were generated using both Gadag's method and the proposed method. Furthermore, the naturalness of the sentences resulted by Gadag's method and the proposed method was compared. Based on Figure 15 it can be concluded that the final score of the paraphrased sentences generated by the proposed method was better than that of the paraphrased sentences generated by Gadag's method. Thus, it was proved that the proposed method could paraphrase sentences well, even when it was implemented on sentences that were not from the corpus.

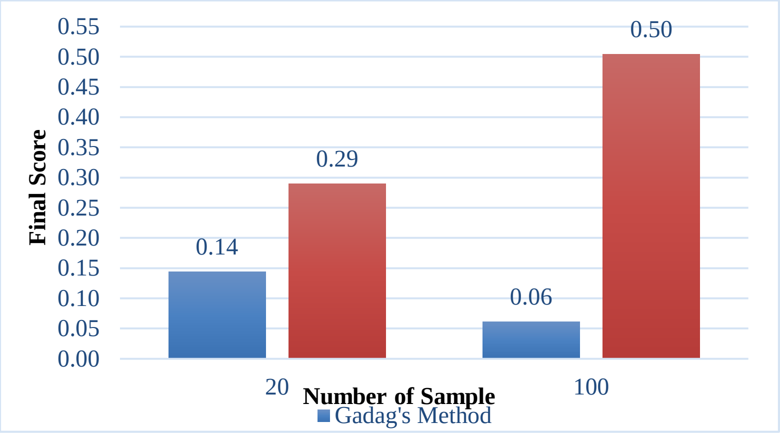


Figure 15 Final score of sentences from the different corpuses.

5.2.2 Result of Experiment Using Human Judgment

Based on the result of the experiment using human judgment represented in Figure 16, it can be concluded that the naturalness of all paraphrased sentences generated by the proposed method was better than by Gadag's method. The reason is the same as given in Subsection 5.2.1.

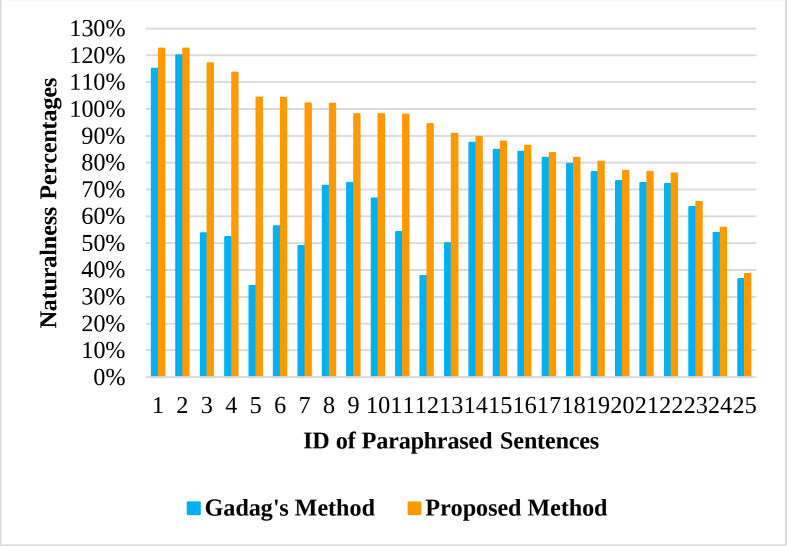


Figure 16 Experimental result for human judgment. Naturalness percentage comparison between Gadag's method and proposed method using questionnaire.

The naturalness percentage could be more than 100% because the paraphrased sentence may be more natural than the original sentence (Twitter). However, several paraphrased sentences had the same naturalness between both methods, for example, sentences 2 and 14-25. The naturalness is the same when there is no ambiguity in the original sentences, such as in sentence 2, '*Diandra dukung pebalap Indonesia di Moto-moto*'. In this case, the context of the paraphrased sentence using Gadag's method and the proposed method are the same, so that the naturalness based on human judgement is close.

5.3 Conclusion

Based on the results from evaluation by human judgment and Meteor Universal Tool, the naturalness of the sentences paraphrased using contextual substitution was better than that obtained by Gadag's method. However, in some cases where there was no ambiguity, the naturalness difference between the paraphrased sentences resulted from Gadag's method and the proposed method was not significant. This is the contribution of the proposed method.

References

[1] Pantel, P. & Lin, D., *Discovery of Inference Rules from Text*, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Data Mining, pp. 323-328, 2001.

[2] Niu, C., Zhou, M., Liu, T., Zhao, S. and Li, S., *Combining Multiple Resources to Improve SMT-based Paraphrasing of the Model*, Proceedings of the 46th Annual Meeting of ACL, 2008.

[3] Ioannis, A.V., Mittal, T.V., Riezler, S. & Liu, Y., *Statistical Machine Translation for Query Expansion in Answer Retrieval*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 464-471, 2007.

[4] Snover, M., Madnani, M., Dorr, B.J. & Schwartz, R., *Terplus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate*, Machine Translation, **23**(2-3), pp. 117-127, 2010.

[5] Durme, B. V., Callison-Burch, C. & Ganitkevitch, J., *PPDB: The Paraphrase Database*, in Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 758-764, 2013.

[6] Gadag, A. I. & Sagar, B.M., *N-gram Based Paraphrase Generator from Large Text Document*, 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 91-94, 2016.

[7] Arman, A.A., Putra, B.A., Purwarianti, A. & Kuspriyanto, *Syntactic Phrase Chunking for Indonesian Language, Proceedings of ICEEI*, pp. 635-640, 2013.

[8] Wicaksono, A.F. & Purwarianti, A., *HMM-based Part-of-speech Tagger for Bahasa Indonesia*, Proceedings of 4th International Malindo Workshop, 2010.

[9] Winstein, K., Tyrannosaurus lex. Open source. Available at http://web.mit.edu/keithw/tlex, 1999.

[10] Chang, C. & Clark, S., *Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method*, *Computational Linguistic*, **40**(2), pp. 403-448, 2014.

[11] Twitter, @tempodotco, 129210 tweets, 01 January 2012-10 June 2016, 10:20 a.m., https://twitter.com/tempodotco.

[12] Twitter, @kompasdotcom, 3410 tweets, 09 February 2012-18 June 2012, 12:32 p.m., https://twitter.com/kompasdotcom.

[13] Twitter, @kompascom, 81759 tweets, 03 March 2015-01 June 2016, 06:54 a.m., https://twitter.com/kompascom.

[14] Denkowski, D. & Lavie, A., Meteor Universal: *Language Specific Translation Evaluation for Any Target Language*, Proceedings of the EACL Workshop on Statistical Machine Translation, 2014.

[15] Lavie, A., Sagae, K. & Jayaraman, S., *The Significance of Recall in Automatic Metrics for MT Evaluation*, Proceedings of Conference of the Association for Machine Transition in the Americas (AMTA), pp. 134-143, 2004.

[16] Banerjee, S and Lavie, A., *Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,* Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, 2005.

[17] Twitter, @Pikiran_rakyat, 76205 tweets, 14 July 2009-17 December 2019, 13:02 p.m., https://twitter.com/pikiran_rakyat.

[18] Twitter, @Jawapos, 47187 tweets, 1 January 2019-17 December 2019, 13:02 p.m., https://twitter.com/jawapos.