OPEN ACCESS
J. OF INISTA, VOL. 1, NO. 1, PP.27-39, SEPT 2018

E-ISSN : 2622-8106
DOI: 10.20895/INISTA.V1I1

Journal of Informatics, Information System, Software Engineering and Applications

Comparing Rule-Based Translation to Syntax Tree Diagram Translation

Heri Heryono

Fakultas Teknik, Teknik Informatika, Universitas Widyatama Bandung
Jl. Cikutra No. 204A Bandung

heri.heryono@widyatama.ac.id

Accepted on September 28, 2018

Abstract

Machine translation has developed more effectively to be specified as modern linguistics characterization. It indicates the formalization of translating process by machine within humanly approaching. Recently, in order to help people communicate indirectly through written, machine translation has its own system to break and match the codes through comprehensive database of words pocket. In linguistics perspective, machine translation always be assisted by its definite performance through derivative computational system. The problem appears when users need to translate ambiguous sentence from English to *Bahasa Indonesia* or *vice versa*. The paper used rule-based system which contains particular parts in order to run the process of translating; starts from tokenization, pre-processing, reordering and morphological process which lead to output accuracy. And used syntactic diagram translation, which was carried out manually by generating sentences into small units. The result of this research showed that RBS runs by standard linguistics system; and the output of translating refer to basic translation without giving any alternative translation nor meaning. While using syntactic diagram, there are possibility to get alternative result not only by linguistics features but also based on signification meaning of both translations. Those techniques run by considering linguistics features in order to help users to get translation output better.

Keywords: machine translation, linguistics, rule-based system, syntactic diagram

I. INTRODUCTION

THE terminology of natural language relates to human language that carried out in conversation in term of conveying information. From this perspective, natural language should be appropriate to be implemented in technology, especially computerized system. Since the system required language to convey output in form of natural language. Computerized things then spread to translation tools that basically essential to help people in understand particular language they do not know. Basically, the translating process requires knowledge and experience; it is not a simple process. Translator and interpreter need simultaneously process to get perfect translation through study. By computing system, the process of translation only carries out several hours to be translator. In the beginning, it became a challenge to create simple process of translating by computerized system. Considering that every language has its own characteristic, its own vocabulary root and its (grammatical) rules and standard that must be processed well to get a precise output.

Dealing with translation process, machine translation (MT) is doing automatic text translating from particular natural language. In other word, natural language will need precious process to be a perfect output that will also understandable by the user. [13] The idea of implementing translation process in computerized system had emerged for a long time ago; it started in 1950s-1980s that was carried out through linguistics study by obtaining information

and features in linguistics to do the process. The previous process was simple, by generating data through dictionary database and grammar rules that would completely take several step process in fixed-and-rigid output only based on dictionary meaning without considering the sense of words when they were translated in different cultural context or even in translating idiomatic phrases that would led to another meaning or connotation. That process called rule-based machine translation (RBMT).

Recent use of machine translation develops through its completion in linguistics aspect and in machine system followed the rules of languages. Machine translation quality is measurable by detecting the detail of the output. Basically, the undetected details of translation lead to improvement to the next system that humanly closer to natural language. The steps of translating process are able to be evaluated by language experts, but recently by the improvement of machine translation, the translator will be positioned as post editor or the third party of the process; since the result or output of the translation has nearly overtaken the perfection. [1]

A tree diagram in linguistics is commonly used in displaying hierarchical structure of a sentence which is generated by sets of rules. [5] Basically, the structure of phrase in grammar is endocentric to constrain possible phrase structure. In modern linguistics, especially in syntax, binary branching such as in tree diagram, becomes a basic principle of merging two categories in grammar that are semantically relevant to each other. [16] It has its own rules in separating sentence based on phrases that lead to exact meaning. In order to get accurate meaning, it might be an alternative to be used in translating process, but it should be manually done by user or they may use some programming methods that have been acquired semantically.

II. LITERATURE REVIEW

Most researches about machine translation relates to NLP, and also it relates to method of translating sentences or phrases that are ambiguous. As it is presented to a statistical phrase-based translation model that uses hierarchical phrases— phrases that contain sub-phrases (Chiang, 2005). In machine translation, the essential feature of its architecture is NLP module which applies Semantic Inferential Model (Kavirajan, 2017). The method of machine translation leads to assembling of part of words that are matched to their lexical meaning in database available in the storage. It turns to be result without giving any alternative translation, while the sentence is ambiguous, means that the sentence will have more than interpretations. Lexical items of second language needs reconfiguration by the learners since they will receive as in the first language. It is essentially different types of lexical items configured in their first language when they only obtain translation from machine translation (Xu, 2007). Applying linguistics part to translate ambiguous sentences/phrases will be more effective in order to get second alternative meaning or interpretation. The system will classify every production in formal language grammar (Kate, 2006).

III. RESEARCH METHOD

The current issue of MT is about its precision in translating words, phrases or sentence that improve well. In case of linguistics perspective, MT applied some features of linguistics since the beginning of its appearance in 1950s. Computational models required natural language as an intermediary to be acquired by users. The process itself run simply by breaking and matching the words from source language to dataset; then they were corresponded to appropriate words in target language dataset.

A. SYNTACTIC AND SEMANTIC AMBIGUITY

It is one of language variation categorized to be interesting yet complex; and it occurs in every language. Ambiguity leads to a way in which words could be selected and/or constructed into particular meaning. The problem relates to partial statement within the language.

In linguistics area, there are several types of ambiguity; words, phrase, clause, syntactic and semantic ambiguity. Since the term of 'pure syntactic ambiguity' emerged as center of ambiguity, it could be defined as ambiguity in which the variant readings of a sentence involve identical lexical units; in this case, the ambiguity is essentially as a matter merely of the way the elements are grouped together." [6]

In this research, there are only one type of ambiguity; semantic ambiguity. It focuses on the fallacy of meaning within one sentence. It appears to be bias when people (reader) tend to get information from this sentence. Especially when people simply translate the sentence using machine translator which has lack of semantically experience or background in converting the information through translating process. It affects to the misleading

information which one person may get different perception, especially when they read a news. Semantic ambiguity can be simply described as when a particular word has two different interpretations. These two different interpretations are two different interpretations, of that one word. At the beginning of a word identification trial, meaning units were placed into a random pattern. By this chance, such a igniting pattern is such to be closer to one of the two meanings of an ambiguous word than to the only meaning of an unambiguous word. [8]

In order to explain about semantic ambiguity, the example below will be shown in combined with translating methods through machine translation and syntax tree diagram method. The sentence will lead to the different translation by these two methods, and it gives the reader more comprehensive thought when the true and logic meaning are acquired through the process.

B. RULE-BASED SYSTEM

Initially, the rule-based system was built on couple of main elements; sets of situational facts and sets of rules of how to reckon with the facts mentioned. In translating, the RBS provides less possibility to set the output, since the set of facts are only provided based on dictionary. It means, RBS has no basic knowledge of data training when it runs to translate. The scheme of translation process in RBS applies linguistics features, such as syntax and semantics. In term of translating, Rule-based system deals with translation rules on linguistics. It concerns to capture syntax and syntactic variation in both translated languages. [7] Basically, in this process, the system receives sentences as an input from the users then continues the process by passing it to the parser, in this term, Stanford parser. It may generate some features that is usually used for translating. Next step is generating the token, or it is called tokenization, POS tagging, dependency of the letter and then syntactic parse tree. Part of Speech (POS) processing will be generated by parser and it will involve morphological rules. When the process passes the morphological rules, the system will continue to check data from database of target language and it performs the results that will be given to the next process in order to generate features that could be possibly used as the output data set. But the process will be corresponding with the availability of dataset in the database. It means that the translation process is running well through sequential process in computer system. After the process has finished, system of morphological generator will provides morphological information for those words that have been translated; the contents will be about case marker, tenses marker, gender (if necessary) and grammatical marker. [10]

It happens to be mistranslated from source language to target language when the input token unavailable in database. The way of machine translation to translate this case may take transliterated from Unicode database; the output will need post editing process manually as consequence. There are four process within the RBS includes sentence simplification, preprocessing and reordering.

A. Sentence Simplification

Machine translation recently accommodate complex sentence which actually as a development of previous machine translation that based on dictionary only. In this process, simplifying a large or/and complex sentences to be partial simple sentence that purposed to sort out sentences and increase the output accuracy. It is the essential feature in MT since the sentences are available to be split out by their main clause, subordinate clause and conjunctions. This process comes to be the early stage of translating process in MT.

B. Preprocessing

This technique runs by converting input data into the process-able structure that the MT can recognize. It may be carried out before the translation process begins. There are several pre-processing steps that are implemented in the system:

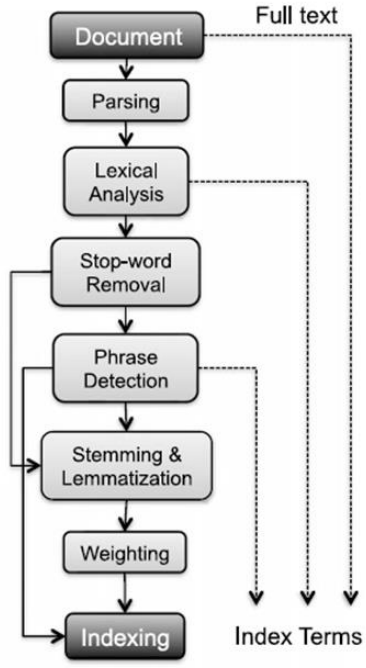


Fig. 1 The pre-processing flowchart

1. Tokenization
   It means that users input sentences by different units of words which is called as *tokens.* For example, the segmentation of words form. Sentences may separate into basic linguistics unit; such as words, phrases, sentences, paragraphs and any other larger/wider form than paragraphs.
   [12] Tokenization relies on separating sentence by space character. This purposes to obtain clear separation that may continue the next process of translating.

2. Filtering (Stop-word Removal)
   It is a continual process from tokenization that takes part in the next translating process. The process of removal word tends to removing words that do not have essential meaning in translating output. In *Bahasa,* the words such as *ini, itu, yang, ke,* etc., will not be included in the next process of translating. [14] In *Bahasa,* there are 720 of stop-words that are listed on a table, then it becomes reference when the system needs some particular words to be used in the next process, the system will check into the table. When it doesn't exist in the table, then the word will be used.

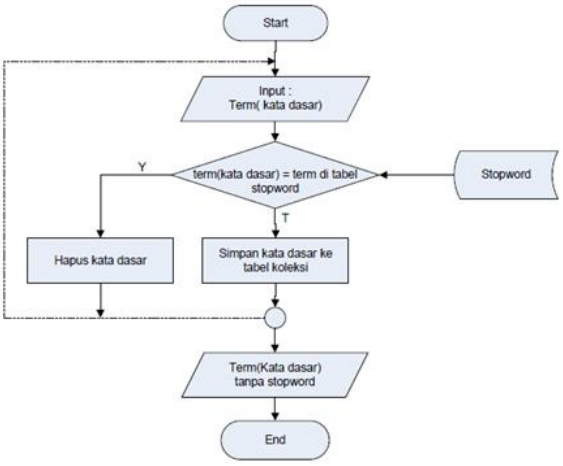


Fig. 2 Flowchart of *Stopward removing* in Bahasa

3. Lemmatization
   In other terminology, lemmatization also called as word stemming, but it doesn't require to produce a stem. It normalizes the affixed-words that need to be formed as basic word or natural word. For example, when it appears the word *working* or *works*, it would be changed to be normal form as *work* that stands as the infinitive base form. [11]

C. Reordering

In some languages, there are several distinguish organizations of words order. Not only in grammatical structure or constrain but also the other convention rules involved in that languages. Related to this word arrangement, machine translation would get fundamental challenges. Typically, the observable of words strain is in noun phrase. [4] In this case, the phrase "an expensive white car" will be translated in *bahasa* as "sebuah mobil putih yang mahal." In MT systems, the process of tokenization, parsing, stop-word removing and lemmatization have been derived from dataset existed in the table that is available in providing the words set. Yet, after pre-processing process, those words then reordering from the target language into source language based on the valid rules of those particular languages. In this case, English to *bahasa*, which has their own rules. In *bahasa*, the order of words will be reversed as English does, such as *white car* changed to *mobil putih.*

In this process of reordering, the system read the possible structure and organization involved in arranging the output based on the target language rules. Reordering process of words or phrases is a step applied in *automatic evaluation* leveling in MT. the result of this process will be implemented in training step (forming of language model and translation model) and also testing level (decoding sentence).

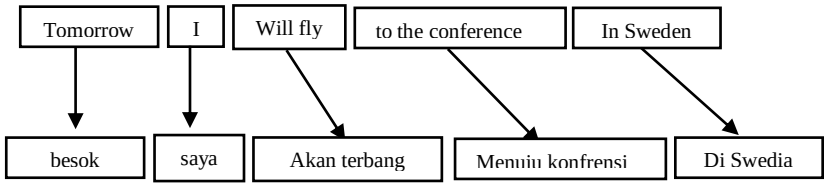


Fig. 3 *phrase reordering model*

C. SYNTAX TREE DIAGRAM

When discussing about morphology, it surely does have hierarchical structure to the way of morphemes put together. Syntax concerns to the way of words are arranged into larger units and it refers to the sentence that usually as a source of analysis of syntax. Therefore, syntax relates to the study of sentence structure. [17] One of the most general methods to show representation of structure of syntax is tree diagrams. This is very common when learner analyses sentence structure in order to get precision model of one whole sentence structure with its parts. Yet, the tree diagram gives some symbols as its representations dealing with labeling the parts of the tree in that hierarchical organization of sentence. There are *det*, *N*, *NP*, *VP*, etc. this tree diagram has definitely worked well to analyze sentence into a breakdown pieces which led to exact separation in obtaining the meaning or context.

Basically MT has its own property to analyze and breakdown the sentence into smaller pieces and make an output as they are combined after particular process within. But, sometimes, machine/computer doesn't have syntactic knowledge as human does. There are at least three syntactic knowledge that are owned by human:

1. Human can produce unlimited sentences and also can create sentence that they never heard before. This could be the characteristic of human mind of creating sentence, it doesn't relate to the meaning or sense of the sentence, but it is about the way human creates the sentence itself.
2. Grammar knowledge can emerge the very long sentence; instead of creating sense or meaning, we prefer to create or compose the long sentence.
3. System translating carried out by human contains a lot of possibility, not about BLEU score of machine translation that reaches 90% but about the syntactic knowledge in arranging the words into the proper meaning and sense.

In order to obtain the closest meaning from source language to target language, KTS *(Kernel's Terminal String)* is usually applied to map natural language to the representative of formal meaning/sense. The classifier in syntax tree diagram, which the output is KTS, refers to parsing method in obtaining the most appropriate meaning in the

particular sentence, especially the sentence which is ambiguous. [9] When certain words are combined in one whole structure, it emerges the hierarchical structure as in morphological structure diagram. The different is that in the top of diagram is sentence, not phrase or word. It tells whole information including part of speech, words order, etc. Syntax actually concerns with the assembling words into phrases and into sentence.

The truth about natural language is that it can be categorized and classified into smaller part or group. It manages the possibility in analyzing structure, categories, part of speech, phrase or even the ambiguous classification happened in that certain sentence as well. Natural language processing considers as a pattern in determining the meaning within sentence input to the machine translation that should be as exact as human translation result. In order to get the most exact meaning from a sentence contained ambiguity, tree diagram would be the proper way to terminate it.

Tree diagram represents the relative of root node and its top of the leaves. There are various representative analysis for the sentence that can be described as phrase marker, parse tree or Kernel's string analysis. Tree diagram gives the visual representation of how sentence should be divided into smaller parts by particular parts that pointed to whole exact meaning of the sentence, especially in translating sentence from source to target language which contains ambiguity. More textbooks perhaps provide tree diagram as the quickest and most efficient representation of an essential hierarchical properties of the proper sentences. [3] In this case, tree diagram would be the most chosen method to describe and show sentence pattern by its parts in order to find out elements of the sentence that make ambiguity to be translated. This method is far conventional than machine translation, yet the result of translation could be the closest one in determining the proper translation (result/output). In tree diagram, there are some symbols that represent syntactic categories in form of abbreviations. Commonly, there are (at least) seven abbreviations that usually involve in tree diagram. [17]
It starts with S represents sentence, NP for noun phrase, VP is verb phrase, PP is prepositional phrase and also the generative parts that will support the tree diagram in determining the sentence parts.

S = sentence
NP = Noun Phrase
N = Noun
VP = Verb phrase
Art = article
V = verb
PP = Prepositional Phrase

All of these symbols sometimes appear in one whole sentence part, but particular sentences do not have all of the parts, as it is shown in the example of simple sentence below:

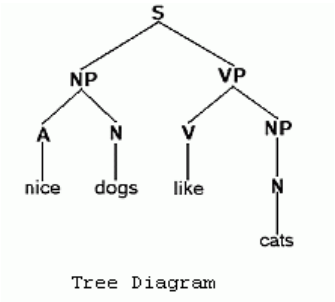


Fig. 4 The tree diagram of Syntax

By translating sentences, the reader expect an exact meaning from the result. But somehow, when translating process is running, machine translator actually works as its database, not the context of the sentence. Using this method, Syntax tree diagram, people have their own decision to determine sentence meaning with more than one option to be chosen as result of translation. We can indicate the parts of the sentences to be translated into their smallest parts. For example, when we examine this sentence:

The tourist saw the astronomer with the telescope.

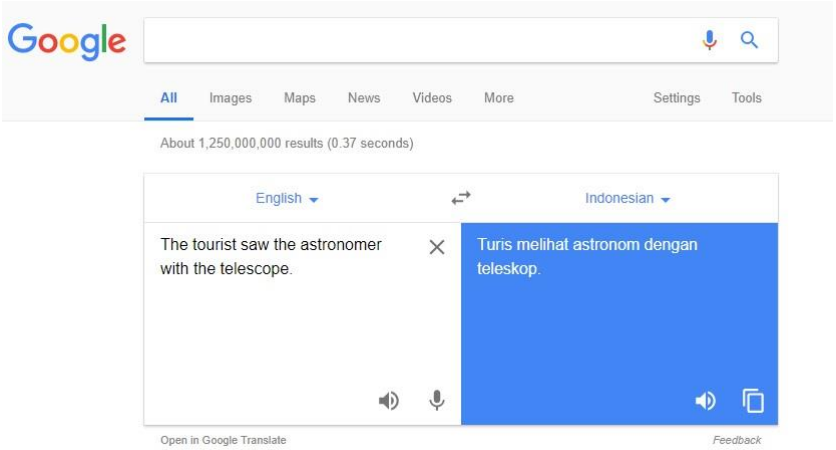


Fig. 5 Translating using MT (Google Translate)

Using only machine translator, you only get one item as the result of translating. The machine will give no more alternative translations as the comparison of whether the sentence ambiguity or not. In this term of ambiguity, machine translation only provides some dataset, no more clue about meaning. But, by using Syntax tree diagram; it will be broken-down into smaller parts and when it contains ambiguous meaning, the diagram also will give alternate version of translation. When it comes to Bahasa Indonesia, people will only receive single result and it will become the only reference. The result itself transforms from database provided by the system that is matched from the input (words) that the user types in the board. By applying tree diagram, the ambiguous sentence will be separated into at least two optional results, as it is shown below by the similar sentence example:

The tourist saw the astronomer with the telescope.

Option 1:

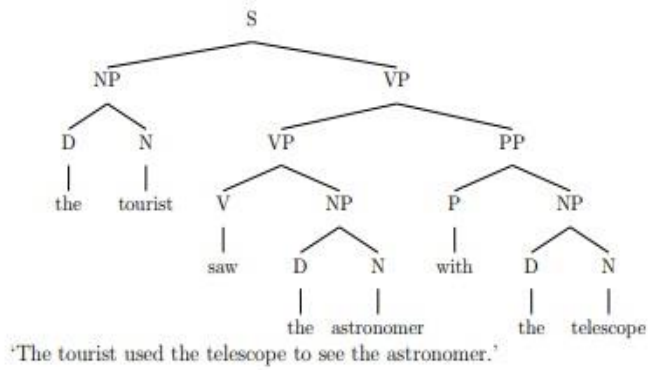


Fig. 6 Syntax tree diagram process

By using tree diagram, we may break down the sentence that:
the tourist as Noun phrase,
saw as main verb,
the astronomer as second noun phrase,
with as preposition and
the telescope as third noun phrase.

In this section, as you can see that the sentence partition has only two main parts. Part one is NP (noun phrase) that only contains determiner and noun and the rest is as part two that has verb phrase and prepositional phrase. In part two, the sentence owns the subsection explaining about what *the tourist* does. By these sectional grouping, the reader may conclude that; *the tourist saw the astronomer* as the main point of the sentence, while *with the telescope* is additional part that is attached as supplementary explanation for *the astronomer*, not for *the tourist.* Hence, this sentence will occur as the exact meaning by analyzing the parts.

Option 2:

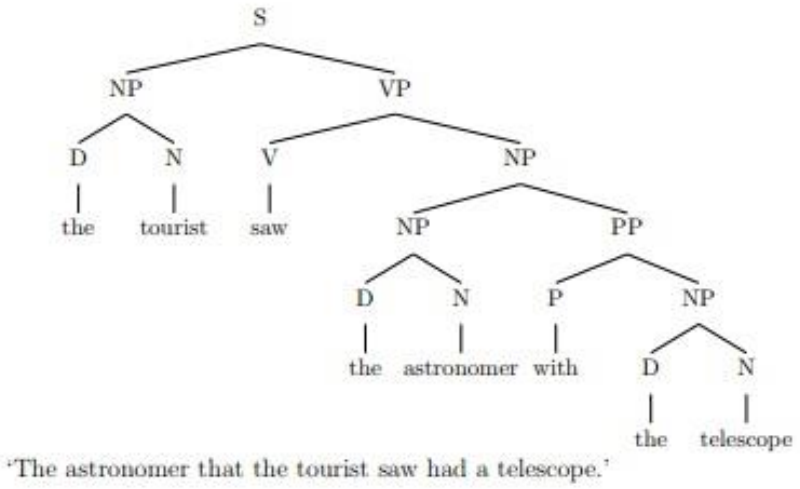


Fig. 7 Syntax tree diagram process

By using tree diagram, we may break down the sentence that:
the tourist as Noun phrase,
saw as main verb,
the astronomer as second noun phrase,
with as preposition and
the telescope as third noun phrase.

In this tree diagram, you will see the distinctive meaning from the first option above. As you can see that the sentence partition has also two main parts. Part one is NP (noun phrase) that only contains determiner and noun and the rest is as part two that has verb phrase and prepositional phrase. In part two, the sentence owns the subsection which is categorized as *noun clause.* The main sentence is only *the tourist saw astronomer.* To get more specific *astronomer*, the sentence then added *with the telescope* to make the second subject (*astronomer)* clearer and definite. By these sectional grouping, the reader may conclude that; *the tourist saw the astronomer* that has *the telescope.* As the result, this sentence will occur as the exact meaning by analyzing the parts.

IV. RESULTS AND DISCUSSION

In some cases, sentences may include several multiple meanings and thus the meanings of these sentences are describe-able in distinctive of word groups or even in dissimilar tree diagram. Here is the example when we use machine translator (in this case is *Google Translate)* in order to translate the sentence contained ambiguity meaning. The research provides some example taken from internet to be discussed and analyzed by those two different method; by machine translation and by Syntax tree diagram.

Here are some headline news title taken from internet as the data source for this section.

1. *They decided on the boat.* [16]
2. *Old men and women.* [11]
3. *Willies wrote a book about Nixon.*[2]

These data sample will be analyzed by two methods, machine translation and Syntax tree diagram.

Data 1:

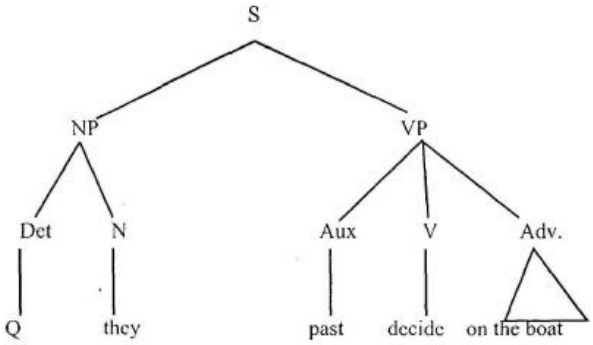
They decided on the boat.

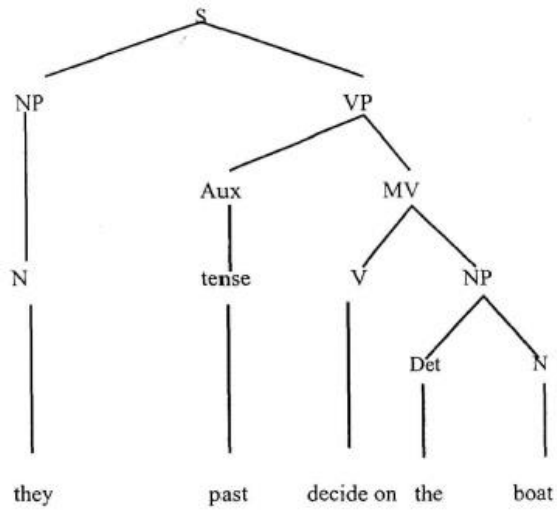
a. Using machine translation

Fig. 8 Translating using MT (Google Translate)

Based on the translation above, taken from translator, the meaning refers to an activity of some people (*they);* thus the sentence has no more explanation more detail to adhere the context which leads to ambiguous meaning. Hence the sentence will be plainly received by the reader as one single meaning sentence and it may probably lead the reader to the miss interpretation situation. Result in bahasa Indoneisa refers to some people who decided to stay in a boat. While, that is the closest interpretation to this sentence, but when it analyzes further, the meaning will lead to another interpretation. Whether *they decided something but the place of deciding was on the boat*, or *they only decided to stay in the boat* and it doesn't relate to another activity.

b. Using Syntax diagram tree





In this sentence, it contains ambiguity since the sentence may have double interpretation. The first interpretation is:
It is obtained when the partition in the tree diagram is:
They decided something by the place is on the boat.
NP = they
Det. = Ø
N = they
VP = decided on the boat
Aux = past
V = decide
Adv. = on the boat

While the second interpretation refers to the activity that they did, in this case, they decided to stay in a place- the boat. The verb *decide* along with *on* which collocates to its function as lexical unit. So, it

functions as transitive which is followed by object. In other hand, the second interpretation leads to a meaning:
They decided to stay on the boat.
The tree diagram parts will be:
NP = they
Det. = Ø
N = they
VP = decided on the boat
Aux = past
MV (main verb) = decide on the boat
V = decide on
NP = the boat
Det. = the
N = boat

Data 2:

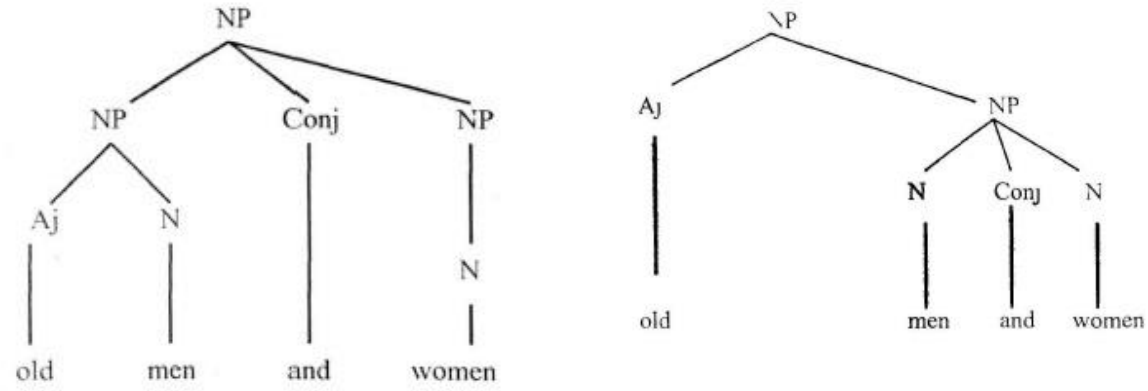
Old men and women

a. Using machine translation

Fig. 9 Translating using MT (Google Translate)

By using machine translator, the output/result for the translating is stuck in one whole package of translation. It may occur a lost in translation condition to the readers when they receive the result. This ambiguous sentence will be able to be reconstructed by its parts manually using tree diagram.

b. Using Syntax diagram tree



As it is shown above, the phrase is basically contains ambiguity. It takes two model of syntax tree diagram to overcome the problem of interpretation; it results two interpretations. The first interpretation is:
The phrase will probably have two main meaning:
NP = old men
Aj. = old
N1 = men
Conj. = and

NP = women
Det = Ø
N = women

This interpretation leads to a meaning of some old men and general ages of women, the definite explanation about the age is only for men, not for women, since the tree diagram only emphasizes the age to the men.

While the second interpretation is refer to the *equality of age* between men and women. So, the tree diagram is set as:

Aj. = Old
NP = women and men
N1 = men
Conj. = and
N2 = women

Data 3:

They cleaned the vase in the store.

a. Using machine translation

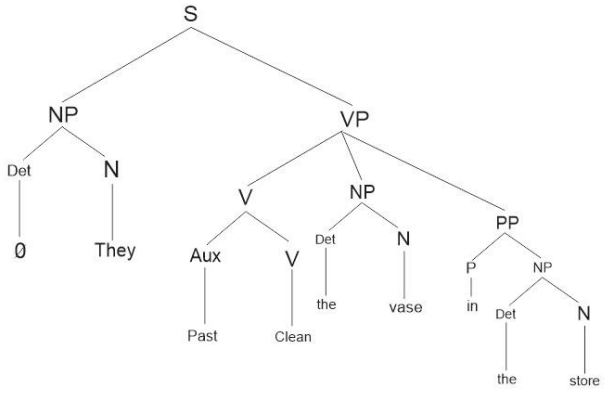
Fig. 10 Translating using MT (Google Translate)

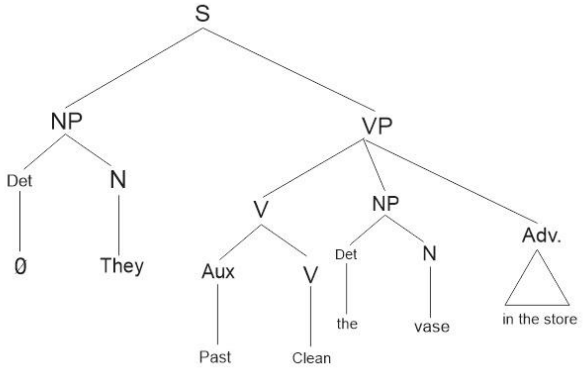
The result shows that the sentence has vogue and ambiguous aspect in its interpretation. It has double interpretations, which can be determined as:

They cleaned the vase that is placed in the store.
They cleaned the vase and they did it in the store.

What can be generated as the exact meaning of the translation if only the reader use machine translator to receive the information? The very best result of this ambiguous sentence could be solved by using tree diagram.

b. Using Syntax diagram tree





The tree diagram above shows the different structure between the interpretations contained in one sentence. By using this syntactical tree diagram, the parts of ambiguous sentence are clearly analyzed and the reader may have alternative representation through the result of translation.

If we breakdown the parts, it will be:

They cleaned the vase that is placed in the store

NP = they
Det. = Ø
N = they
VP = cleaned the vase in the store
Aux = past
V = clean
NP = the vase
Det = the
N = vase
Adv. = in the store

They cleaned the vase and they did it in the store.

NP = they
Det. = Ø
N = they
VP = cleaned the vase
Aux = past
V = clean
NP = the vase
Det = the
N = vase
PP = in the store
P = in
NP = the store
Det = the
N = store

Based on those analysis, it results the alternative interpretation for one ambiguous sentence; and it takes two different tree diagrams of syntax to get those alternative interpretations.

V. Conclusion

From the analysis carried on by the author in translating method of ambiguous phrase and sentences, it could be concluded that using syntax tree diagram was more effective than machine translation applying rule based system. By using tree diagram, the author found that those sentences and phrase might have alternative interpretation that differently emerged to dissimilar result. Ruled based machine translation conveyed information (translation) only based the database they kept in system. It matched and proceeded the string of words in sentences or phrases without giving any alternative results of translations, in this case was ambiguous sentences/phrases. While using syntax tree diagram, the structure broke down the parts of sentences/phrases into smaller unit within their segmentation. The result was that the translation occurred in any alternative meaning or interpretation based on the structure they made.

ACKNOWLEDGMENT

I would like to thank the all colleagues and also library media specialists for supporting this research. And also to Widyatama University which facilitates this research, and I also grateful to the members of the department for their patience and support in overcoming numerous obstacles I have been facing through my research.

REFERENCES

[1] A. Burchardt, K. Harris, G. Rehm, and H. Uszkoreit. *Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation. –In: Proceedings of the LREC 2016 Workshop "Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem".* Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (LREC-2016), located at International Co, 2016.

[2] Akmajian, Adrian, Richard A. Demers, and Robert M. Harnish. 1979. *Linguistics, an introduction to language and communication*. Cambridge, Mass: MIT Press.

[3] Baker, L. (1998). *English Syntax*. London: The MIT Press

[4] Chiang, David. 2005. A hierarchical phrase-based model for statistical machine transla- tion. In Proceedings of ACL

[5] Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics (sixth edition)*. Oxford: Blackwell

[6] D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press

[7] G. Sangavi, K. Mrinalini, P. Vijayalakshmi, "*Analysis on Bilingual Machine Translation Systems for English and Tamil"*, International Conference on Computation of Power, Energy Information and Communication (ICCPEIC), 2016.

[8] Joordens, S., & Besner, D. (1992). Priming effects that span an intervening unrelated word: Implications for models of memory representation and retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 483-491.

[9] Kate, J,. Rohit., Mooney, J,. Raymond. *Using String-Kernels for Learning Semantic Parser.* In Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), pp. 913-920, Sydney, Australia, July 2006.

[10] Kavirajan, B., Kumar, Anan. *Improving the Rules based Machine Translation System using Sentence Simplification (English to Tamil).* ICACI, September 2017.

[11] Plisson, Joel., Lavrac, Nad., Mladenic, Dunja. *A Rule based Approach to Word Lemmatization.* Department of Knowledge Technologies Jožef Stefan Institute.

[12] Robert A. Hall, Jr., *An Essay on Language.* Philadelphia: 1962.

[13] Russell, S. & Norvig, P. *Artificial intelligence: a modern approach*. Pearson Education, Ltd. 2004.

[14] Shao, Yan. *Tokenization and Sentence Segmentation.* Department of Linguistics and Philology, Uppsala University, 29 March 2017.

[15] Tala, F. Z. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam the Netherlands.

[16] Xu, J. (2007). *A Comparative Study of Categories in Generative Grammar and Parts of Speech in Traditional Grammar*. Research on Chinese as a Second Language, 3.

[17] Yule, G. (2010). *The Study of Language*. Cambridge: Cambridge University Press

[18] West, C. and Zimmerman, D. (1983) "small insults: a study of interruptions in crosssex conversations between unacquainted persons'. In Thorne, B., Kramarae, C. And Henley, N (eds) (1983) Language, Gender and Society. Rowley: Newbury House. 103- 17.