Collocational Equivalence in Machine Translation[2]

Alvin Taufik
Universitas Bunda Mulia
ataufik@bundamulia.ac.id

Abstract

The objective of this research is to discover the potentiality of machine translation, in this case represented by *Google Translate* software, in creating equivalent for the collocations in Bahasa Indonesia as the Source Language into English. The equivalence which becomes the focus in this paper is based on the concept of equivalence above word level as proposed by Mona Baker (1992). The types of collocations refer to classifications by Imran, et.al. (2009). The research is focused on the results of the machines' translation of specific texts which correspond to the limitation as stated previously, and its concordance to the commonly accepted usage as demonstrated in the Corpus of Contemporary American English (COCA). As a result, it is found that the translation made by Google Translate has a high frequency of occurrence and naturalness.

Keywords: *Collocation, Machine Translation, COCA*

1. Introduction

The Machine Translation (MT) which is available nowadays was seriously developed initially in the 1950s. The first big budget research on MT was done jointly between Georgetown University and IBM. Initially, the research was not developed using sound linguistic approach. Instead it was approached in two different ways, contrastive and empirical trial and error. Only later on, some rare linguistic related approaches, prominently proposed by Zellig Harris and Noam Chomsky, are referred in the project to develop fully automated translation. In addition, in the beginning, this Machine Translation, or in non English speaking countries often referred to as 'Automatic Translation (AT)', was focused on scientific and technical documents (Hutchins, 1995).

Although at that time hopes were high with the development of formal linguistics, as well as the development of computing, a very critical argument made by Bar-Hillel (1960) pointed out that it is impossible for an automated translation to break through semantic barrier; thus creating a fully functioning MT. Moreover, in the later years, the sponsorship for a fully automated translation was postphoned and more research was later focused on machine to aid in translation. This is because the development project of machine translation was calculated to be more expensive than human translation, and it was considered that there was no immediate need for an automated translation. The

[2] This journal entry was presented as a scientific paper for the 1st International Translation & Interpretation Symposium with the same title and by the same researcher.

research on MT went quiet for a decade. It was later reignited with the installation of Systrans.

Fast forward to the 1980, IBM has once again created a software which was later developed into what is known today as *Google Translate*. Bellos (2012) mentioned that unlike earlier AT, *Google Translate* (GT) no longer deals with meaning. Instead of taking the language as something which needs to be deciphered using artificial intelligence, it uses statistical methods to find the most probable acceptable language pair from the previously submitted documents. In other word, it relies heavily on a corpus of data. It is exactly for this reason that, in this research, the result of the GT later on will be paired with the corpus available for public, namely Corpus of Contemporary American English (COCA).

Why choose collocation? Why collocation in Bahasa Indonesia? Collocation in definition is the tendency of a number of words to **repeatedly** coexist in one utterance. The key word is in the word repetition, as this can be loosely interpreted as some word pairs has high frequency of coexisting with each other in different situation of utterance. This can relate to the way GT works at the moment, as GT is also relying heavily on statistical frequency of coexistence of word pairs. Thus is the reason for choosing collocations.

As for the reason of why Bahasa Indonesia is chosen, there are two main factors behind it. The first is concerning the number of research on collocation in Bahasa Indonesia. Imran et. al. (2009) stated that there is still too little research concerning collocation in Bahasa Indonesia. So this research is hopefully will provide additional information on collocation in Bahasa Indonesia in general, and specifically in their translation. The other factor, or reason, is related with the Target Language (TL) comparison choices. To identify the accuracy and naturalness of the translation results, there should be a standardized corpus of data to which the results can be compared. If English is the Source Language (SL) it will be difficult to compare the result as there is no definitive corpus of Bahasa Indonesia currently. On the other hand, COCA is readily available in public domain.

As stated above, in contrast to the past mechanism, GT has shifted its focus from meaning based into statistical search of compatibility probability between language pairs. It needs to be seen whether this new method adapted by GT is proven to be effective. This research is aimed to find out whether accuracy and naturalness has been achieved by GT in translating, especially in translating language pairs such as collocations. Therefore, the research question can be formulated into:

- How accurate and natural Google Translate is as compared to the data found in Corpus of Contemporary American English?

There are at least two variables in this research which needs to be clarified first before the analysis is started. The first point is on collocation itself. Baker (1992) identifies collocational equivalence (it was collocated with 'equivalence' since the focus is on its translation) as equivalence which goes beyond words (equivalence above word level). In her book, there are interesting points regarding collocation. The first is about collocational range and markedness. She mentioned that some words have broader range than the others; some words have more words to collocate with compared to others. Range of collocation is determined by two main factors: specificity and number of senses. In terms of specificity, the more specific a word is the lesser the words collocate with it. In terms of senses of the collocations, it means that words which have more senses, such as the word 'run' which can have the sense of 'manage' if collocated with 'business', and has the sense of 'operate' if collocated with 'service', the broader the range of the collocation is.

The pattern of collocation created by the range has been known and often is identified as one's 'linguistic repertoire'. Yet, collocation keeps on growing, sometimes to create new image. These new, and sometimes, unfamiliar ones according to Baker are called marked collocation. However, it is also possible that collocations are unfamiliar because they belong to specific registers.

In translation, collocations have created problems. One problem is when the translation of collocation focuses more on the source language. Another problem is the common misinterpretation of collocation in Source Language. This happens usually because the words involved in the Source Language collocation is familiar to be collocated with another word in the Target Language. Other problem is related to accuracy and naturalness. Sometimes, when a translator is trying to be accurate it disregards naturalness, and vice versa. This often occurs in translation of other linguistic items, including collocations. This issue on accuracy and naturalness is also one main point to be proven in the translation of collocation using Google Translate. Like other items in a language, collocation can also be culture-specific. This too has created problems in translation. This issue is also related to the previous point on naturalness and accuracy.

The previous was brief explanations on collocation and its problem when concerned with translation. The second variable which is equally important concerns with the types of collocation which exist in one language. Since the Target language will be

compared with existing database, the Source Language will refer to some explanations written by Imran et.al. (2009). In definition, their concept on collocation is similar to those which have been theorized in regards to English collocations. They divided collocations into two, grammatical and lexical collocations. They further classified them into collocations with unique sense, common sense, and specific sense. Notice the similarity in the theory since Imran et.al. also used Baker as one of their theoretical basis. In addition to their polar categorization, they also classify lexical collocation into collocations which is formed of Nouns, Adjectives, and Verbs, whereas grammatical collocations include 'functional' words.

2. Research Methodology

This research is qualitative in nature. The research aims to describe the 'degree' of accuracy and naturalness of the translation by Google Translate. The analysis of the data is done by comparing the results of the translation made by Google Translate with the corpus of data available in the COCA. The SL data is gathered from random source which includes collocation as stated by Imran et.al. (2009). In addition, all SL included in the analysis will be added with the context in which such collocations are found. This inclusion of context serves as the basis for the comparison of naturalness between SL and TL. For this research, the analysis will only be focused on equivalence of lexical collocations.

3. Findings

Imran et.al. (2009) stated that lexical collocations can have the patterns of the following:

Table 1. Patterns of Lexical Collocation (adapted from Imran et.al., 2009)

Types	Pattern	Examples
L1	Noun + Verb	Air mengalir, petir menggelegar
L2	Noun + Adjective	Kopi pahit, teh kental, gerak lambat
L3	Noun + Noun	Es batu, kopi susu, hujan batu
L4	Verb + Noun	Membajak sawah, mengemudikan mobil, naik pangkat
L5	Adjective + Verb	Cepat sembuh, lambat mendarat, berani bertanggungjawab
L6	Verb + Adjective	Lari cepat, berpikir logis, jalan santai, bicara tinggi
L7	Noun + Adverb	Tahun lalu, tahun depan, halaman belakang
L8	Verb + Adverb	Berlayar langsung,

L9	Adjective + Noun (specific meanings)	Sakit hati, keras kepala, besar mulut, rendah hati

And, here are the results of the translation for these groups of collocations

Table 2. Indonesian and English Collocations in Comparison

Type	SL (Bahasa Indonesia)	TL (English Translation)	COCA Result
1.	**Air mengalir** Context: '..**Air mengalir**karena adanya perbedaan ketinggian..'	**Running water** Context: 'Rinse the leek well, flipping layers under **running water'**	1309 in frequency, number 1 on the list
2.	**Kopi pahit** Context: 'Bagi penggemar **kopi pahit**, rasa pahit dan aroma dari kopi tersebut tentu akan memberikan suatu kenikmatan yang khas.'	**Bitter coffee** Context: 'He had sipped a cup of **bitter coffee**.'	28 in frequency, number 34 on the list
3.	**Es batu** Context: **'Es batu** memiliki manfaat bagi kecantikan kulit Anda.'	**Ice cube** Context: 'Just like if you take an **ice cube** out of the freezer...'	488 in frequency, number 4 and 5 on the list
4.	**Naik pangkat** Context: '25 Perwira Tinggi TNI **Naik Pangkat** KBRN'	**Move up** Context: '..when they took a test to **move up** the promotion ladder'	1128 in frequency, number 4 on the list
5.	**Cepat sembuh** Context: 'Temanku, Semoga **Cepat Sembuh**.'	**Speedy recovery** Context: 'She grimaced when the anchor and crew wished her a **speedy recovery**.'	67 in frequency, number 2 on the list
6.	**Jalan santai** Context: 'Agar tetap bugar, **jalan santai** sejauh 2,5 km sehari saja dapat mengurangi risiko terkena penyakit jantung'	**Leisurely stroll** Context: 'I slowed down and began a **leisurely stroll** through many of Messier's masterpiece.'	40 in frequency, number 2 on the list
7.	**Tahun depan** Context: 'Isu perkara masuknya Honda	**Next year** Context: 'The Pentagon budget will	16523 in frequency, number 1 on the

	menjadi sponsor utama Moto GP **tahun depan** sangat kuat.'	shrink slightly **next year**.'	list
8.	**Berlayar langsung** Context: 'Karena angin sedang baik, diputuskan untuk **berlayar langsung** ke Cina.'	**Sailing directly** Context: 'Indus traders began **sailing directly** to Arabia.'	4 in frequency, number 23 on the list
9.	**Besar mulut** Context: 'Karena **besar mulutnya** sehingga banyak mulut yang membicarakannya.'	**Vain glorious** Context: 'He shuts his ears to them and imagines, instead, talking with the **vain glorious** old explorer whose tales left him feeling lost, and full of questions.'	-

4. Conclusion

As can be seen above, the result of Google Translate in the translation of different types of lexical collocation has proven to be accurate and natural. This can be seen from the frequency concerning those specific collocations, the non-literal translation of the collocations (e.g. in collocations such as *es batu*, *jalan santai*, and the collocation with specific meanings) and its position on the list of collocates.

One thing to be considered, however, is on the context of the Target Language. As an example, in the phrase 'running water' although there are context which are similar to those in the SL, most of the contexts of the phrase 'running water' are actually related to water supply in one's house. The same is also observable in the TL of *naik pangkat*, which is 'move up'. In its case, move up are mostly related to moving from someone's house according to COCA. Yet, like stated previously, there are also some contexts in which the phrase 'move up' is used in a similar context as in its SL.

References

Baker, M. (1997). *In Other Words: A Course book on Translation*. London: Routledge.

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers* 1, p. 91-163.

Bellos, D. (2012). *Is That a Fish in Your Ear?: Translation and the Meaning of Everything*. Straus and Giroux: Faber & Faber

Corpus of Contemporary American English (COCA).

Hutchins, W.J. (2001). Machine Translation over Fifty Years. *Histoire, Epistemologie, Language*, Tome XXII, fasc. 1 (2001), p.7-31.

Imran, I., Said, M., and Setiarini, N.L.P. (2009). *Kolokasi Bahasa Indonesia*. Proceedings from PESAT. Depok: Universitas Gunadarma.