

Semantic-based Malay-English Translation using N-Gram Model

Nooraini Yusoff, Zulikha Jamaludin, Muhammad Hilmi Yusoff

School of Computing, UUM College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia.
nooraini@uum.edu.my

Abstract—Most of the existing machine translations are based on word-for-word translation. The major obstacle in developing such a system is natural language is not free from ambiguity problems. One word may have more than one semantic, and vice versa. Herein, we propose a semantic-based Malay-English translation using an n-gram model. The Malay-English translation is not a word-for-word basis but is dependent on the semantic meaning of the Malay phrase. In particular, a bigram is used to approximate the probability of a word by using the conditional probability of the preceding word. For this study, whenever the semantic ambiguity occurs, the English word with the highest probability value is chosen to translate the Malay word (or 2-sequence Malaysia word). The proposed technique has been tested with three categories of sentences namely easy, moderate and complex. The performance of the proposed Malay-English translation is based on human judgement that demonstrates an averaged validity ratio of positive value. The positive value indicates that at least half of the respondents agreed that the translation outputs are at least “still make sense semantically”. The contribution of the proposed method can be ascribed to the enhancement of word-for-word translation for solving the ambiguity issue in Malay-English translation.

Index Terms—Machine Translation; Malay-English Translation; N-Gram; Semantic; Ambiguous.

I. INTRODUCTION

The Malay language is widely used as an official language in Malaysia. Many important texts and documents in either public or private sectors are written in the Malay language. On the other hand, English language is widely used everywhere in the world. At the same time, Malaysia has different kind of people from different kind of cultures. These people have adopted English as their second language and used the language in their official work. Besides that, Malaysia is going to become a hub in education and business in the region. Thus, having an automated system that can translate Malay into English texts is very desirable. However, the major obstacle in developing such a system is natural language is not free from ambiguity problems. One word may have more than one semantic, and vice versa. This is not a problem in the Malay language only, but also in other natural languages. Therefore, to develop a system that can make a translation like a human does is not an easy task.

The high development of machine translation (MT) systems from English to other languages is currently being the famous issue rather than translation from other languages into English. Although machine translation has been introduced in past 30

years but MT system from Malay language to English is not yet in rise. Historically, the study of Malay language in MT has been conducted since 1984 with the establishment of Unit Terjemahan Melalui Komputer (UTMK) at Universiti Sains Malaysia [1], but this is not a point to put Malay to English translation on top. Until now, there are only a few MT systems concentrating on translating Malay to English. Three of them are Citcat Sdn. Bhd. (www.citcat.com), Google Translator (translate.google.com), and UTMK [2].

Although these MT systems are extensively used by many people, some for commercial purposes, they are still not free from flaws. For example, some of the translated sentences lost their meaning due to the restructuring in the target sentences. The worse translation may happen if the source language includes affixes and words with multiple meaning (ambiguous words). Google translator is fast and easy to use, and it claims to provide adequate general content translation for more than 50 languages. However, due to its limitations, it could create a false sense of security because the meaning is not detected or conveyed accurately. For example, given a Malay sentence

“Makcik ke pasar malam menaiki kenderaan awam.”,

yielding the following result:

“Aunt night market ride public transportation.”

The result is worse for Citcat

“Auntie to night market travel by public transport”.

Other than the MTs from Malay to English, the MTs from other languages to English may also face their own difficulties. For example, translating bidirectional English/Persian would have problems like natural language ambiguities, anaphora resolution, processing idioms and slangs [3].

This study focuses on designing techniques to identify and resolve ambiguity problems in Malay texts. To the best of our knowledge there is no research thus far has used semantic-based knowledge in conducting Malay-English automated machine translation. We put forward the claim that semantic-based translation is the best translation tool, if doable, because it will give the correct meaning instead of lateral meaning which could be meaningless in terms of words associations and the semantic equivalents when deriving meaning from the

translations results.

By having a reliable computational tool for the more mundane, time-consuming tasks such as MT, much of the time of a human translator is no more wasted in manual lexicographic searches, and in document editing and formatting. Time consuming as they may be, these are the simplest tasks that a translator must perform, and therefore the easiest to automate effectively. Further, the following are some of other direct contributions of this study.

- The method of extracting useful information from bilingual corpus with a correct semantic can be used by other researchers and scientists alike to improve future algorithm for improving the semantic as well as the grammar in the target language, thus to enhance the quality of the translation.
- The prototype produced could encourage Malaysian to use MT facilities that will eventually improve their knowledge in Bahasa Malaysia and English so that language barrier can be decreased.
- An easy to use tool with a readable translation will especially be useful to the Malay speaking users in understanding English conversation at workplaces. Companies and institution that are looking for a translation tool (Malay-English) may thus benefit from this work.

This paper is organised as follow: Section II discusses the background and related works. Experimental methodology and results of our proposed Malay-English machine translation are discussed in Section III and IV, respectively. The paper is concluded in Section V.

II. AMBIGUITY ISSUES IN MACHINE TRANSLATION

Handling ambiguous sentences is one of the key issues in MT. Translation is said to be ambiguous when a word in a source query may have more than one sense. However, this crisis usually exists in any translation process. "Ambiguity is a linguistic feature" [4], and there are numerous types of ambiguity in the study of machine translation, information retrieval, grammatical analysis, speech processing as well as text processing. Eberle [5] has found one of the hard problems after doing his first steps towards finding ways for translating text automatically that is ambiguity of words and structures. Translation ambiguity is not only a hard problem but also a basic problem to be resolved [6]. A string with multiple interpretations is also declared as ambiguous. Previous studies illustrated that there are differences between resolving ambiguity between two possible meanings of a word, and ambiguity between two possible interpretations of a phrase. Translation ambiguity can be as many as five, six, or even more possible translations. Such ambiguity creates a major challenge in real-life bilingual language processing.

Semantic ambiguity is a part of specification of the grammar of a language where the most semantically ambiguous sentences are not noticed by listeners but typically discovered only by linguistic research [8-9]. In MT, semantic ambiguity could be a resultant of lexical ambiguity, anaphoric ambiguity or syntactic ambiguity.

Lexical ambiguity occurs when a word has multiple

meanings. For example, "*Saya rasa takut*", the word "*rasa*" in Bahasa Melayu could be translated as "*to taste*" or "*to feel*". Meanwhile, anaphoric ambiguity occurs when a phrase or word refers to something previously mentioned, but there is more than one possibility. Let a prime phrase in Malay "*Azimah mengajak Salmah makan malam*", followed by a later phrase "*tetapi dia alergi kepada makanan laut*". To whom the word "*dia*" (in English "*she*") refers to, is ambiguous, that one might be asking, who is allergic to sea food? Furthermore, a paper of Proverb Treatment in Malay-English MT [1] gave some examples of semantic ambiguous in Malay proverbs. "*mata air*" can be lover, or underground water resource and "*air muka*" can be face, or pride. A sentence like "*I feel blue*" should be translated as "*Saya berasa sunyi*" where "*blue*" in the sentence is not a kind of colour but a feeling (lonely) [2]. Complexity takes place when the MT need to know the definite meaning of proverbs or words [1].

Semantic ambiguity in a sentence is not only caused by the multiple senses of meaning, but also the syntactic structure of the sentence. Syntactic ambiguity arises from the association between the words and clauses of a sentence, and the sentence structure implied thereby. For example, a sentence "*Azimah makan roti bersama keju yang dibeli dari Tesco pada setiap pagi*" is ambiguous, as "*setiap pagi*" can be conjoined with "*makan*" or "*dibeli*", and "*dibeli*" can also be conjoined with "*roti*" or "*keju*".

A sentence could also be regarded as genuinely ambiguous in its semantic if the sentence really can have two different meanings to an intelligent hearer (i.e. human). In such cases, the translation is too tightly dependent on the context. Hence, the disambiguation process might require more complex analysis and mapping of the domain knowledge, and at some point this rather impossible to be done by a machine. For brevity, consider a sentence "*Ahmad dan Salmah sudah berkahwin*". The sentence has an ambiguity - is it they married to each other or both married to different persons? The semantic of the sentence could only be accurately defined if both the source and target persons share the context, or there is a prime sentence that could help to hint the semantic. For example, if the given sentence is preceded with a sentence "*Salmah adalah tunang kepada Ahmad*", then the semantic of the sentence is clear. To handle the ambiguity problems, word sense disambiguation plays its role as to identify the correct sense of each source word [3,8,9].

Syntactic ambiguity differs with the lexical and anaphoric ambiguities where it arises from the location of the words in sentences, not from the range meaning of a single word. In other words, the sentence may be interpreted in more than one way due to ambiguous sentence structure [10]. Meanwhile, lexical ambiguity is a "pervasive problem in natural language processing" [11]. It is both very common and very difficult to clear up if to compare with a sentence being syntactically ambiguous. In the Lexical Ambiguity and Information Retrieval project by Hussein and Bahareh [4], they took semantic and syntactic ambiguity as two types of lexical ambiguity. They conducted a few experiments to get a better understanding of lexical ambiguity and its effect on information retrieval. The result showed that lexical ambiguity is not a significant problem in documents containing large number of words in common with a query.

III. METHODS

In this section, we present the steps involved in developing a semantic based Malay-English translation. We begin with analyzing the input texts. Then we discuss in details on the implementation of n-gram in solving the translation ambiguity issues. Next, we explain how the accuracy of the translation output is measured.

A. Analyse Input Texts

Sentences were extracted from the Traveller's Guide Book Malay to English [12]. The extracted sentences were treated as test cases in this study. We chose to use Tourist's dialogue as our test case because Malaysia has become one of the tourism destinations in the world and English language is very important to use in speaking with the foreigners or when we become a traveler to other countries.

Each sentence structure is analysed in term of its part of speech and categorised according to simple, moderate and complex sentence levels. Later, the sentences are parsed by extracting the corpus content from Excel-Format. Then the sentences are further extracted into words with their parts of speech, synonym list and meaning list (Malay- English) and words' dictionary by using natural language processing functionalities in Python. All the extracted sentences and words are stored post-wise in an nltk-phyton texts database.

In analysing the input texts, a corpus and dictionary are needed. Corpus selection is to select the input texts within the scope, while dictionary contains Malay to English words translation. The purpose of analysing the input texts is to categorize ambiguous words or phrases in Malay texts. We only aimed at ambiguous Malay words –those that may have more than one translation in English. For instance, '*selamat*' is translated as '*safe, good, secure*'. Our targeted deliverable from this analysing process is a list of ambiguous Malay words or phrases.

B. Semantic-based Translation using N-Grams

In this step, we develop a method consisting of techniques to translate the semantics from Malay texts into a target language (i.e. English). This include attaching a correct semantic to the ambiguous words in Malay texts and translating the words into relevant English meaning.

The Malay-English translation is not a word-to-word basis but is dependent on the contextual (i.e. semantic) meaning of the Malay phrase. In this study, ambiguity in a translation occurs whenever there exists a Malay word that has more than one meaning in English. For example, the word "*adik*" in Malay could refer to "*sister*" or "*brother*" in English. The Malay-English machine translation process involves sentence parsing, ambiguous words identification, and semantic translation.

a. Ambiguous Words Identification

After we parse a Malay sentence into words, we list all the single and two-word sequence words made up the given sentence. This to assume a direct translation could be done with a maximum of two Malay words. From the corpus, we initially augment each sentence with a special symbol <s> and </s>, at the beginning and end of the sentence, respectively.

For example, when prompted with a sentence "*selamat pagi encik*", the semantic translation engine parses the sentence into the following single and two-word sequence words:

[selamat, selamat pagi, pagi, pagi encik, encik]

Then the meaning for each word (and two-word sequence) is looked up from the dictionary. As a result, the English translation for the sentence is shown as in Table 1.

Table 1
Malay-English Translation Result for a Malay Sentence "selamat pagi encik"

Malay	English
selamat	[safe, good, secure]
selamat pagi	null
pagi	[morning]
pagi encik	null
encik	[sir]

In the translation process, a two-word (i.e. a bigram) English translation is given a priority to be chosen as the best translation prior to identifying the ambiguous words. This to say, for example, whenever "*thank you*" is found to be an English translation (not described in Table 1), that would have higher priority compared to other possible meaning(s). In this study, we define Malay ambiguous words as words with more than one meaning in English, e.g. "*selamat*" in Table 1.

b. Semantic Translation

After we identify a Malay sentence consisting of words with more than one meaning in English, we then predict the appropriate semantic translation for those ambiguous words. For this purpose, we use a language model, n-gram, that predicts the next word from the previous n-1 word. An n-gram is an n-token sequence of words: a 2-gram is commonly called a bigram, is a two-word sequence of words, e.g. "*please accept*", "*accept my*", or "*my apology*", and a 3-gram is called a trigram, is a three-word sequence of words, e.g. "*please accept my*" or "*accept my apology*". Using the n-gram language model, computing the probability of the next word turns out to be closely related to computing the probability of a sequence of words.

In our implementation, for ambiguous words found in a Malay sentence, we predict the semantic translation of the words using a bigram model. This is to assume that, the ambiguity of meaning of a Malay word could be resolved by predicting the meaning of two-word sequence consisting of the word followed by a later word. For example, a Malay word "*terima kasih*", depending on the application, perhaps the appropriate semantic translation is "*thank you*" instead of "*accept love*".

The bigram model approximates the probability of a word given all the previous words by using the conditional probability of the preceding word. Therefore, the probability of a word depends only on the previous word (i.e. Markov assumption).

Here we use the simplest way to estimate the probability that is by using maximum likelihood estimation (MLE). To compute a particular bigram probability of a word w_n given a previous word w_{n-1} , we compute the count of the bigram $C(w_{n-1}, w_n)$.

w_n), and normalised by the sum of all the bigrams that share the same first word w_n . Hence, $P(w_n|w_{n-1}) = C(w_{n-1}, w_n) / C(w_{n-1})$.

For brevity, consider a mini-corpus consisting of 3 sentences as follows:

```
<s> I am Ahmad </s>
<s> Ahmad I am </s>
<s> I do not like green eggs and chicken </s>
```

The bigram probability for “I” is calculated as:

$$P(I|<s>) = 2/3 = 0.67$$

in which, from the corpus, there are two bigrams where “I” is found given a previous <s>. In other words, I is to be the start of a sentence (“I am Ahmad” and “I do not like green eggs and chicken”). The calculations for some of the bigram probabilities from the corpus are:

$$\begin{aligned} P(\text{Ahmad}|<s>) &= 1/3 = 0.33 \\ P(\text{am}|I) &= 2/3 = 0.67 \\ P(<s>|\text{Ahmad}) &= 1/2 = 0.5 \\ P(\text{Ahmad}|\text{am}) &= 1/2 = 0.5 \\ P(\text{do}|I) &= 1/3 = 0.33 \end{aligned}$$

In our implementation, we predict the translation of a Malay word with more than one semantic meaning in English based on the highest probability of a bigram composed of each of the possible English translations and given the preceded meaning. An example of a Malay-English translation is illustrated and described in Figure 1.

For the implementation of the proposed n-gram based language model, we develop a prototype of Malay-English semantic translation. Java is the core language to programme the engine, and we deploy Java-based Android for the user interface.

C. Evaluate the Accuracy of the Translation

To validate the functionality and robustness of the proposed Malay-English translation method, we probed the trained corpus with three sets of sentences according to the level of difficulty (i.e. complexity of sentence) namely easy, moderate and difficulty. For each level, we tested on sentences without and with semantic ambiguities from the trained corpus, incomplete sentences and unseen sentences.

In general, our validation method on semantic accuracy takes into accounts the humans’ judgements, who consider the vocabulary, part of speech, and positions of the target words. The subjects are Malay native speakers (aged in between 30-45 years old), whose English is their second language.

Prefaced by detailed instructions, the subjects’ answers are collected via a question posed after every resulting translation. A total of five subjects are adequate to judge on each translation [13, 14]. The formula is followed from the content validity ratio (CVR) measurement [15] since the aim is at what it superficially appears to measure. Such validity requires the experts to evaluate whether the output is semantically similar to the input. Lawshe [15] proposed that each of the experts respond to the question in the form of 'essential,'

'useful, but not essential,' or 'not necessary'. We adapted these answers into “semantically correct”, “still make sense semantically”, “totally wrong semantically”.

Lawshe claimed that if more than half of the panellists indicate positive answers, then we can claim the answer is valid. The formula is written as (1):

$$CVR = \frac{(n_e - N/2)}{(N/2)} \quad (1)$$

where CVR = content validity ratio,

n_e = number of experts indicating at least "still make sense semantically", i.e. 1,

N = total number of experts (in our case, 5).

This formula yields values which range from +1 to -1; positive values indicate that at least half of the experts rated the item as at least the translated sentence “still make sense semantically”. The mean CVR across items may be used as an indicator of overall correctness.

IV. RESULTS AND DISCUSSION

For our MT prototype, we have selected nine cases to test the performance of our proposed Malay-English translation method. The nine cases comprised of both ambiguous and non-ambiguous words chosen to represent different levels of sentence complexity. The test cases are comprised of easy, moderate and difficult sentences. The performance of the translation engine is measured based on the work by Lawshe [15].

A. Corpus Training

We have trained an English corpus related to a Malay corpus under study. For the developed prototype, we extracted and trained nine English sentences to test the functionality of the proposed method described in Section III. As the result of corpus training, the translation prototype listed all the bigram probabilities, $P(w_n/w_{n-1})$ from the trained English corpus. Some examples of the output are as follows:

```
P(malay|am) = .0312
P(i|think) = 1.0000
P(brown|mrs) = 1.0000
P(fine|am) = .3438
P(sit|please) = .5000
```

The probability value, $P(w_n/w_{n-1})$, indicates the popularity of usage in the corpus. Hence, the higher the value would lead greater chance for a particular bigram to be selected whenever semantic ambiguity exists.

B. Testing

During the testing, we probed the prototype with a Malay sentence to be translated into English. For example:

```
Please enter a Malay sentence to be translated
into English:
>> Selamat pagi encik
```

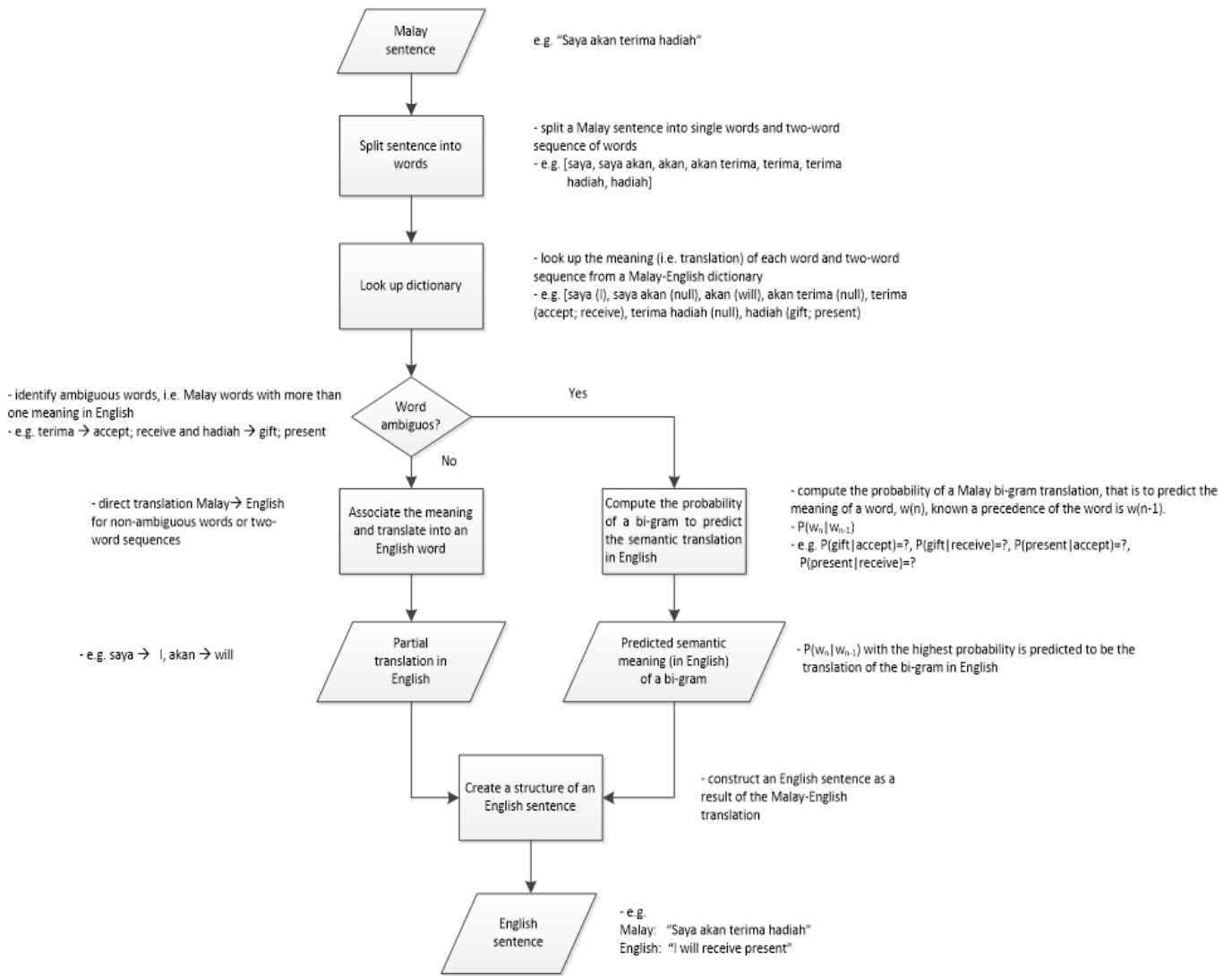


Figure 1: Flow of a Malay-English semantic translation using bigram

The translation engine then parsed the sentence into a list of single words and two-word sequences.

```
>> [selamat, selamat pagi, pagi, pagi encik, encik]
```

After that, all possible translations for each word were given. At this stage, the engine was to identify the semantic ambiguities in which there were particular Malay words with more than one semantic translation. The output of this stage is as follows:

```
selamat: [safe, good, secure]←semantic ambiguity
selamat pagi: null
pagi: [morning]
pagi encik: null
encik: [sir]
```

The ambiguity is resolved by choosing the highest value of the associated bigram probabilities, i.e. $P(\text{morning}|\text{safe})$, $P(\text{morning}|\text{good})$ and $P(\text{morning}|\text{secure})$.

$$P(\text{morning}|\text{safe}) = .0000$$

$$P(\text{morning}|\text{good}) = .5926$$

$$P(\text{morning}|\text{secure}) = .0000$$

The result has shown that the word “morning” preceded by “good” was the most popularly used in the corpus, and hence “selamat pagi” was translated to “good morning”. The semantic ambiguity was resolved as “selamat pagi” could be directly translated from the English dictionary (“selamat pagi: null”).

TRANSLATED SENTENCE:
good morning sir

C. Categorizing the Results

Next we tested the proposed bigram-based semantic translation on a set of easy, moderate and difficult sentences. We follow the content validity ratio, CVR, measurement as proposed in [15]. For our study, the positive values of CVR indicate that at least half of the experts rated the translated sentences as at least “still make sense semantically” (see

Section III.C). Here we highlight some examples of the translation results for easy, moderate and difficult sentences in Tables 2 - 4, respectively.

An easy sentence contains only one independent clause. Basically the translation will go correctly because the sentence is short. In our case, 2 out of 3 test cases for easy sentences indicate positive CVR.

The only one with -ve CVR (third row in Table 2), was due to “*temujanji*” is actually read as “*temu janji*” and looked up as *temu* → “*come together*” and *promise* → “*janji*”. Perhaps our respondents could not perceive this as semantically correct. The NULL output is produced when a Malay word is not registered in English translation.

As shown in Table 3, a moderate sentence contains two or more independent clauses. The sentence is longer compared to easy category and translation using machine is quite challenging. Similarly, (as found in easy test cases) for moderate sentences, 2 out of 3 test cases are rated at least “still make sense semantically” by at least half of the respondents. A large corpus is needed to calculate the probability of words and two-word sequence, so that the translation will be better.

A difficult or complex sentence defined in this study is a sentence that contains one or more independent clauses and one or more dependent clauses. The level of translation complexity is increased from easy to difficult. Nevertheless, to our surprise, all test cases are rated with at least “still make sense semantically” by at least half of our respondents.

The results indicate some potential application of our proposed method in Malay-English translation. We have demonstrated the suitability of n-gram model in MT application in conditions from easy to complex sentences.

D. The Proposed Method versus Google Translate (GT)

The main idea of the proposed method and Google Translate (GT) is to do translation. However, the proposed method scope goes to Malay-English translation only; while GT has two ways translation for over 50 languages. Since GT is very popular among internet users and available as a free online application, the corpus must be very large compared to our proposed method where the corpus is limited to only 2422

sentences. On the other hand, the special part of the proposed method is that the prototype is able to display the sentence structure of translated sentence in form of tree (Figure 2) whenever the translation is true, while GT is not able to do so. Nevertheless, the translation using GT is quite good in speed. The significance of displaying the tree in proposed method is to provide an easy platform to users to understand the rules of sentence structure.

English sentence: “Don't worry. I'll examine you.”

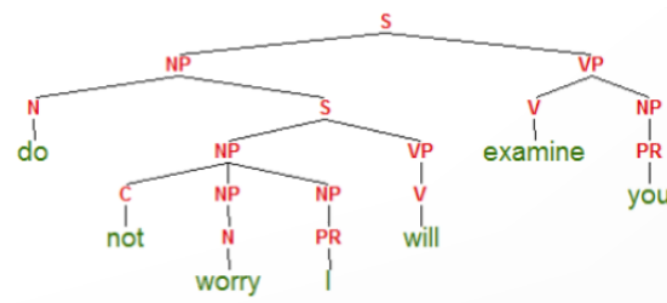


Figure 2: An example of a sentence structure

Although GT is widely used, it somehow has its limitation where it may not detect linguistic or convey it accurately. Hence, GT is improved from time to time. Interestingly, we initially used GT to translate a Malay sentence, “*Makanannya boleh tahan juga tetapi terlalu pedas*”. Firstly, GT translated the sentence as “*The food is okay too but too spicy*”. The output was not as good as the later translated sentence “*The food is not bad, but too spicy*”

Other than that, GT may also misinterpret the grammar of complex structure that may lead to not accurate and precise output [6]. The proposed method may face the same problem as GT since its corpus is limited.

Table 2
Examples of Translation Results for Easy Sentences

Malay Sentence	English Sentence	Machine Translation	Validity Ratio
Lebih kurang sepuluh tahun	For about ten years.	more less ten year	1
Berapakah umur Encik Kassim sekarang?	How old are you now?	how old sir NULL now	0.6
Apakah Encik sudah membuat temujanji?	Do you have an appointment?	what sir already do come together promise	-1

Table 3
Examples of Translation Results for Moderate Sentences

Malay Sentence	English Sentence	Machine Translation	Validity Ratio
Jenama yang sudah terkenal tentulah mahal sedikit daripada jenama yang belum terkenal.	A well-known brand is slightly more expensive than an unknown one.	brand which is already famous definitely expensive a little from brand which is not yet famous	1
Di Malaysia hanya ada dua musim, iaitu musim hujan dan musim panas.	In Malaysia, there are only two seasons, the rainy season and the dry season.	in malaysia only have two season NULL season rain and season warm	1
Selesai saja majlis perkahwinan, mereka terus terbang ke Hawaii untuk berbulan madu.	Immediately after the ceremony, they flew to Hawaii for their honeymoon.	finished only party marriage they straight fly to hawaii for moon honey	0.2

Table 4
Examples of Translation Results for Difficult Sentences

Malay Sentence	English Sentence	Machine Translation	Validity Ratio
Jangan khuatir, saya akan periksa Tuan.	Don't worry. I'll examine you.	do not afraid i will investigation sir	1
Makanannya boleh tahan juga tetapi terlalu pedas	The food was so-so but too spicy.	his food can endure also but too hot	0.6
Tahukah Encik Lim di mana saya boleh membeli peti televisyen yang bagus?	Do you know where I can get a good television set, Mr. Lim?	know sir NULL in where i can buy case television which is fine	-1

V. CONCLUSION

The general context of our work is the treatment on semantic extraction from Malay, specifically on the ambiguous sentences. The resulting translation, in English, remains the meaning in terms of its semantic primes. Although for some translated sentences, the English grammar are incorrect, but the 'conceptual grammar' is retained. The resulting sentences retain combinatorial properties by virtue of the particular concept it represents. The word order and some other syntactic properties could differ from the 100% correct translation, but the underlying combinatorial properties in the translated sentences of the target language are left undisturbed.

As an additional way to confirm that the semantic are retained in the target language, we also generate a visualization of the resulting tree structure for each translated sentence. The tree structure, deduced from parse/semantic tree with logical form features, is hoped to further clarify the meaning of the source sentence being translated.

We should think of ambiguity as a matter of degree, rather than an all-or-none state. Because a word that is unambiguous (consist of only one meaning) still rely on context. In this paper we have decided that there are three types of ambiguity namely pure, lexical and anaphoric. The type of ambiguity that is solved here is of lexical, i.e. the ambiguity that is caused by multiple meanings of a word.

The proposed n-gram method is considered as a major contribution to the field of MT, specifically Malay-English semantic translation. The resulting performance percentage should indicate the suitability of the method used in which the evaluation depends not only on human judgment but also the word class (part-of speech) similarity measures.

ACKNOWLEDGMENT

This research has been funded by the Ministry of Education (Malaysia) under the Fundamental Research Grant Scheme.

REFERENCES

- [1] K. Abd. Rahman, and M. N. Norita, "Proverb Treatment in Malay-English Machine Translation," *Proceedings of the 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013)*, pp. 4-8, 2013.
- [2] Y. Muhamad Nor, Z. Jamaludin, and S. Jusoh, "A Retrospective View of the Promise of Machine Translation for Bahasa Melayu-English," *Found in Translation International Conference*, Universiti Malaya, 2010.
- [3] H. H. Chen, G. W. Bian, and W. C. Lin "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval," in *Proc. 37th Annual Meeting of the Association for Computational Linguistics and Chinese Language Processing*, pp. 215-222, 1999.
- [4] V. D. Hossein, and Z. Bahareh. "A Semantic Study of the Translation of Homonymous Terms in Sacred Texts: the Qur'an in Focus," *Journal of Language & Translation*, vol. 10, no. 1, pp. 45-79, 2009.
- [5] K. Eberle, Semantic issues in Machine Translation, in C. Maienborn, K. von Heusinger, and P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, de Gruyter, Berlin (Band 3), 2012.
- [6] F. Butler. Machine versus Human: Will Google Translate Replace Professional Translators? pp. 1-10, 2011.
- [7] M. Poesio, "Semantic Ambiguity and Perceived Ambiguity", in K. van Deemter and S. Peters, Eds. Cambridge: Cambridge University Press.
- [8] A. Prior, S. Wintner, B. Macwhinney, and A. Lavie, "Translation ambiguity in and out of context," *Applied Psycholinguistics*, vol. 32, pp. 93-111, 2011.
- [9] L. Morhben, A. Zouaghi, and M. Zrigui, "Lexical Disambiguation of Arabic Language: An Experimental Study," *Polibits*, vol. 46, pp. 49-54, 2012.
- [10] I. S. Bajwa, M. Lee, and B. Bordbar, "Resolving Syntactic Ambiguities in Natural Language Specification of Constraints," in G. Alexander, Ed. in *Proc. 13th International Conference Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, vol. 7181, pp. 178-187, 2012, Springer-Verlag Berlin, Heidelberg.
- [11] R. Krovetz, and W. B. Croft, "Lexical Ambiguity and Information Retrieval," *ACM Transactions on Information Systems*, vol. 10, no. 2, pp. 1-32, 1992.
- [12] L. Y. Fang. *Speak standard Malay a beginner's guide*. Singapore: Marshall Cavendish, 2006.
- [13] W. Lowe, and K. Benoit, "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark," *Political Analysis*, vol. 21, no. 3, pp. 298-313, 2012.
- [14] M. R. Steenbergen, and G. Marks, "Evaluating expert judgments," *The Netherlands European Journal of Political Research*, vol. 46, pp. 347-366, 2007.
- [15] C.H. Lawshe, "A quantitative approach to content validity," *Personnel Psychology*, vol. 28, pp. 563-575, 1975.