A Retrospective View of the Promise of Machine Translation for Bahasa Melayu-English

Yusnita Muhamad Noor
Zulikha Jamaludin
Shaidah Jusoh

College of Arts and Science
Universiti Utara Malaysia
06010 Sintok Kedah, Malaysia

ABSTRACT

Research and development activities for machine translation systems from English language to others are more progressive than vice versa. It has been more than 30 years since the machine translation was introduced and yet a Malay language or Bahasa Melayu (BM) to English machine translation engine is not available. Consequently, many translation systems have been developed for the world's top 10 languages in terms of native speakers, but none for BM, although the language is used by more than 200 million speakers around the world. This paper attempts to seek possible reasons as why such situation occurs. A summative overview to show progress, challenges as well as future works on MT is presented. Issues faced by researchers and system developers in modeling and developing a machine translation engine are also discussed. The study of the previous translation systems (from other languages to English) reveals that the accuracy level can be achieved up to 85 %. The figure suggests that the translation system is not reliable if it is to be utilized in a serious translation activity. The most prominent difficulties are the complexity of grammar rules and ambiguity problems of the source language. Thus, we hypothesize that the inclusion of 'semantic' property in the translation rules may produce a better quality BM-English MT engine.

Keywords

Machine translation, Bahasa Melayu-English translation, semantic property

1.0 Introduction

Due to a different communication culture in 1940s, especially in the business sector, a proposal to develop a machine translation (MT) was put up by Weaver (1946). However, for the first two decades after that, not much improvement can be seen. Most machine translation 'systems' were disappointing in terms of their performance and output quality. The failure was due to the complexity of the specific dictionary-driven rules for syntactic ordering, which enable the system to analyze the structure of the syntax. Since then, numerous projects were inspired by linguists and computer scientists. They finally encountered the "semantic barriers" as the problem for which they saw no straightforward solution. Soon, the term semantic-based translation was popular.

The semantic-based MT was introduced in the 1970s. Many MT researches were conducted, and many systems were designed, focusing on various issues in semantic-based translation. Japanese and English were among the most languages used at that time. Only in the 1980s, researchers began to include languages such as Chinese, Japanese, and German. BM however, was not taken into account even though it is one of the world's top 10 languages. The progress has been rather slow and only in early 2000 did MT become an important field in Natural Language Processing (NLP) that attracted many researchers worldwide. The demand for MT kept increasing due to the advancement in linguistics, computer hardware and software technology.

In section 2 this paper reviews the progress of machine translation and the challenges faced by researchers. The focus is mainly on progress of machine translation that takes semantics into account in translating other languages, especially BM into English. Some recommendations for future work is also included in section 3.

2.0 MT and Semantic-based MT

This section describes the progress of MT and reviews the semantic-based MT to English language. The progress of BM to English translation in both methods, manual and automated system, is specifically focused. Some popular and successful systems is particularly analysed in order to decide which systems and methods can be further investigated for implementation in BM-English automated translation.

2.1 The Progress of Machine Translation

A joint project by Georgetown University and the International Business Machines Corporation (IBM) in 1954 has successfully run Russian to English experimental MT system. Public demonstration was held on 7 January 1954 to translate 60 sentences taken from the field of chemistry to introduce the MT to the community. The demonstration has attracted many countries such as in United States, Russia, and Western Europe to get involved in MT. The eagerness was however weaken due to poor output quality (in terms of correct translation). The performance of the MT was too disappointing. The major problems identified at that time were the high number of synonyms in the dictionary and high cost to develop the system (Ornstein, 1955).

In 1964, Automatic Language Processing Advisory Committee (ALPAC) was established in the United States and in 1966 a report was produced by ALPAC on the progress of MT. It concluded that MT was too slow, had a poor output quality, and was more expensive than a human translator. It suggested that an automatic dictionary was a solution to help the human translator and there was no need for further investment in MT (Hutchins & Somers, 1992).

MT started to bloom again in the 1980s when computer speeds and processing power became cheaper. Further advances in linguistics, computer hardware and software spurred research in the field of MT, with Japan taking the lead (Hutchins, 1995). Another push factor was the introduction of statistical and example-based methods for MT. When those two methods came into force, many translation software were developed offering a wide range of language for use by the translator and general public (Hutchins, 2005). However, those two methods have drawbacks. Both approaches lack syntactic and semantic rules in the system. Nearly all operational systems developed at that time depend heavily on post-editing to produce acceptable translations because the system was not capable to translate accurately due to the problem of understanding the semantic structure and some ambiguity rules. With such performance, semantic understanding is still a major research focus in MT.

2.2 Semantic-Based Machine Translation Projects

The meaning of 'semantic' in linguistics perspective refers to the study of how language (sentences/words) conveys meaning. While in the computer science perspective, it is regarded as the purpose of function or program in an application. In laymen terms, semantic means the meaning of words/sentences. Semantic-based translation would then means the translation that can give words connotation as how it meant in the context of the sentence, as oppose to syntax. A semantic-based MT should firstly be able to identify if a sentence has a correct structure. Secondly, it should be able to translate the sentence into a correct structure and meaning. Thirdly, a semantic-based MT should be able to identify a semantically insensible sentence.

A sentence such as "I feel blue" is of a correct sentence structure. It should be translated into "Saya berasa sunyi", i.e. *blue* in the context of the sentence is not a kind of colour. Rather it is a kind of feeling (lonely). Another problem, sentences can be grammatically correct but make no sensible semantic. A famous example is a sentence created by Chomsky (1957) "Colourless green ideas sleep furiously". The sentence is correct in terms of syntax and sentence structure (Figure 1) even though no one can figure out what it means.

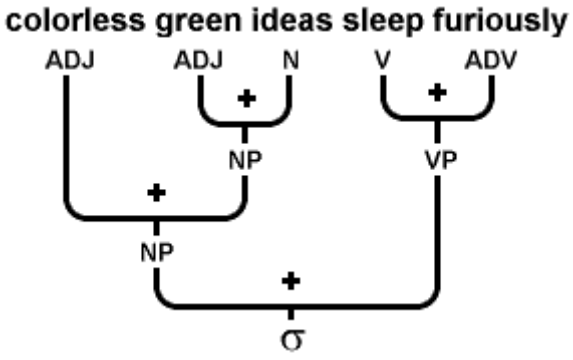

Figure 1: A sentence with correct syntax and structure.
(http://www.knowledgerush.com/kr/encyclopedia/Phrase_structure_rules/)

Since 1975, a number of semantic-based MT systems were developed. Table 1 summarizes 40 of them. Some (40%) take English as the source language, 33% consider English as a target language, while another 27% are geared to other languages.

Table 1: Summary of semantic-based MT

Year	**Language**	**Software**
1975	Chinese-English	CULT
1976	English-French	METEO
1978 & 1986	Russian-German, French-German, English-German, Esperanto-German, German into English and French, Danish-German and Dutch-German	SUSY
1980	Russian-French	ARIANE-78,
1982	German-English, English-German, English-Vietnamese and English-Farsi	LOGOS
1982	Dutch, English and Spanish	Rosetta
1984	Japanese-German	SEMSYN
1985	German-English	METAL
1985	Japanese-English	ATLAS II
1985	French-German	ASCOF
1985	Danish, Dutch, German, English, French, Italian, and later also Greek, Spanish and Portuguese	EUROTRA
1985	Spanish and French, and translate into English	MOPTRANS
1986	Japanese-English	LUTE
1986	Japanese-English	MU
1986	Japanese-German	SEMSYN
1987	Japanese-English	ALT-J/E
1987	Japanese-English	JETR
1987	Japanese-English/English-Japanese	PIVOT
1987	Czech-Russian	RUSLAN
1988	English-French	Critter
1988	Korean-Japanese	NARA
1989	Japanese-English / English-Japanese	AS-TRANSAC
1989	English-Chinese	JFY-IV
1989	English-German	LMT
1990	English-Chinese	TRANSTAR
1991	English-Japanese, French and German	KANT
1992	English, Spanish, and German	UNITRAN
1993	Japanese-English	MMT
1993	Spanish and Japanese	Murasa.ki
1994	Japanese-Russian	JaRAP
1994	German-English	KIT-FAST
1994	English-Japanese	LogoVista E

		to J
1995	Japanese-English	JICST
1999	Japanese-Malay	ALT-J/M
2000	English-Arabic	SEATS
2001	Japanese-Chinese	ALT-J/C
2001	Korean-English	CCLINC
2003	English-Polish	KANTOO
2005	English-Indian	AnglaBharti

		II
2009	English-Persian/Persian-English	PEnTrans

English as target language

English as source language

From the first time MT was introduced, MT that took semantics into account clearly had a better translation quality (in terms of correct translation percentage) compared to non-semantic-based MT systems. The METAL system, for example, a semantic-based MT system, can achieve up to 85% accuracy level (Bennett & Slocum, 1985). The accuracy level achieved by semantic-based MT thus far is above 90%, by PEnTrans MT system (2009), which translates English into Persian.

Studies on semantic understanding were done by the researchers who noticed semantic barriers as a factor for the difficulty in interpreting syntax correctly. Other problems faced are poor English and occurrences of syntactic structures unknown to the parser. These were two problems faced by METEO (1976), whereas incorrect grammar rules was faced by SUSY (1978), which in both cases created ambiguity problems for translation. These difficulties were minimized with the application of better approaches and methods for identifying semantics in sentences as proven by CCLINC (2001) and PEnTrans(2009).

2.2.1 English as a Target Language

Among the languages involved in semantic-based MT are Chinese, German, Japanese, Spanish, Korean and Persian. In the world's top 10 languages in terms of number of native speakers, BM is ranked fifth with a native speaker population exceeding 200 million (*Majlis Antarabangsa Bahasa Melayu* –MABM-- 2008).

Unfortunately, only a few studies have been conducted in translating BM to English. Table 2 provides a list of 11 semantic-based MT systems that translate other language into English.

Table 2: Summary of Semantic-based MT into English

System	Description	Problems/Challenges faced	Accuracy level
CULT (1975)	To translate two Chinese scientific journals, Acta Mathematica Senica, and Acta Physica Senica (Loh & Kong, 1978).	Problem to input Chinese characters.	-
METAL (1985)	Technical translation domain. Required post-editing for high quality output (Bennett & Slocum, 1985).	Difficult to handle technical text of operation and maintenance manuals.	85% correctness of full sentences, experimenting with 1000 pages.
ATLAS II (1985)	Translations for creating English computer manuals (Sato, 1989).	The processing time in translation. The time taken is proportionate to the length of the sentences.	80% correctness in translating automobile service manuals after a joint project with Fujitsu.
MOPTRANS (1985)	To read newspapers on the topics related to terrorism (Hutchins, 1986).	Difficult to handle large scale of texts because it cannot determine conjoined texts.	-
MU (1986)	To translate scientific and engineering papers between Japanese and English.	Many sentences are difficult to understand by native speakers related with the input abstracts in term of its construction and idiosyncratic (Nagao, Tsujii, & Nakamura, 1985; Tsujii, 1987).	For part 1 the higher score is 32.7% and part 2 was 33.3 %.
ALT-J/E (1987)	Automatic Language Translator-Japanese to English, with no pre-editing and pre-writing.	To improve the translation rates of long sentences which (30 words and above) To improve output quality (Ikehara, Shirai, Yokoo, & Nakaiwa, 1991).	The rating ratio for blind and window test was over 60%. The parsing ratio achieved was 80%.
PIVOT	Japanese-English	To analyse the information	-

(1987)	English-Japanese.	structure and to deal with pragmatic issues (Muraki, 1987).	
MMT (1993)	Multilingual MT system.	Word sense selection and the rule-based approach to the disambiguation of natural language - difficult in focus and in handling grammar rulea (Yasuhara, 1993).	-
JICST (1995)	Translate scientific and technical documents -- available for PC and Mac versions (Ashizaki, 1995; O'Neill-Brown, 1996).	Strive to improve the translation quality.	-
CCLINC (2001)	Translate Korean newspaper articles and chemical biological warfare in real time with a large sentence volume (Lee, Yi, Seneff, & Weinstein, 2001).	-	50% correctness, tested with 1600 sentences.
PEnTrans (2009)	English into Persian (PEnT1) Persian into English (PEnT2).	Problem occurred when translating Persian language due the ambiguities arising from the general text (Saedi, Shamsfard, Motazedi, 2009).	Above 90% for PEnT1 in terms of grammatical correctness and 85% completely similar with human translation for PEnT2.

2.2.2 Challenges Faced by Researchers

Challenges faced by the developer in 1970s were the ambiguity in sentences, technical restrictions, and lack of hardware facilities. In the 1980s, hardware facilities were no longer a paramount issue. However, certain problems persisted: those rooted in controlling the grammar rules, eliminating sentence ambiguity, getting information for system implementation, reducing processing time in translating, and the problem to assign an explicit structure to the grammar especially in a situation where large grammars have to be written. On top of that, post-editing still needed to be done in each translation—of all systems (Lau, 1987).

The problem in handling grammar rules was also faced by certain system produced in 1990s, especially in translating Japanese or Chinese into English as experienced in MMT system (Yasuhara, 1993). Systems in the 2000s still cannot resolve ambiguity issues due to the difficulties of the language processing itself, as demonstrated by PEnTrans system (Saedi *et al*., 2009).

2.3 Translation in Malaysia

In Malaysia, manual translation is more active than automated translation. The first MT center, called Unit Terjemahan Melalui Komputer (UTMK), was established in 1980s at Universiti Sains Malaysia (USM). It is a joint project with University of Grenoble of France to develop English to BM MT system. Universiti Teknologi Malaysia (UTM) is another active player in MT. In 1981, UTM established a MT centre and conducted a KANTA project in collaboration with Japan, Thailand, Indonesia, People's Republic of China and Malaysia. The purpose was to produce an inter-lingual MT system among the national languages of the five countries involved. UTM also has developed an English-BM MT system, a joint research project with University of Manchester Institute of Science and Technology (Ahmad Zaki, 1993). In 2002, MIMOS in collaboration with USM developed an English-BM MT system that was claimed to have achieved moderate quality of translation accuracy (Suhaimi, Noorhayati, Hafizullah, & Abdul Wahab, 2006). In late 2006, USM managed to complete an English-BM MT system in collaboration with various parties that uses a large bilingual knowledge bank or BKB (Lim, Ye, Lim, & Tang, 2007).

In 1993, the Malaysian National Institute of Translation (*Institut Terjemahan Negara* or ITMN) was established in Kuala Lumpur, Malaysia, and is responsible for managing the translation for the government. ITMN has developed a machine-aided translation system to assist human translator in translating books from different sources. Again, the main problem that arose was related to ambiguity and different sentence structures that lead to complex grammar rules. Thus, the translation required a pre- and post-editing which entails a longer time and higher cost if judged against using a human translator. Translation in Malaysia remains manual, and sometimes assisted by the use of electronic translation tools such as an online dictionary (Ahmad Zaki, 1993).

2.3.1 Current BM-English MT systems

Currently, there are three translation engines enabling translation from BM to English. The engines are provided for commercial use by Citcat Sdn. Bhd. (www.citcat.com), Google translator (translate.google.com) and UTMK. The obvious drawback of those engines are that the translated sentences lost their grammatical structure and syntax because it changes the arrangement of the translated text. The result is worse if the source language includes affixes and words with multiple meaning. The wrong syntax and grammar structure certainly lead to erroneous translation. Thus, post-editing is unavoidable. Table 3 illustrates some translation results from Citcat Sdn. Bhd. and Google translator.

Table 3: Some translation examples from BM to English

Input sentence	**Output from Citcat.com**	**Output from Google translate**	**The correct sentences**
Saudara fikir Peter ada wangkah untuk dipinjamkan kepada saya?	Relative think Peter exists wangkah to be seconded me?	Brother Peter is thought to wangkah loaned to me?	Do you think that Peter has money to lend to me?
Kami sangat memerlukan tenaga pakar seperti tuan.	We badly needed specialists masterfully.	We really need experts such as master.	We really need an expert like you.
Saudari apa Khabar?	You how are you?	Saudari apa khabar?	How are you?

There are many ambiguous BM words sourced from the instability between syntax and semantics of the language. In order to solve the ambiguity problems, semantic understanding should be applied.

It can be assumed that the success of semantic-based MT depends on (1) the elimination of ambiguity in the source language; and (2) methods to simplify the complex grammar rules. Solving those two upshots would enable better modeling and implementation of the translation engine. One of the possible techniques that can be used to solve the ambiguity and grammar rules problems is by modeling the semantic of a particular source language. A good model will help MT developers to better understand the semantic of a language. It is noteworthy that semantics is language-dependent, so the model should be tailored as closely as possible to fit the source language.

3.0 Conclusion and Future Work

The progress of MT and semantic-based MT (SMT) has been discussed in order to compare their effectiveness and to gauge the challenges faced by MT researchers and developers. Focus here was on MT systems with BI as the target language. This paper has shown that many MT systems were disappointing in terms of their output quality although the field has been studied since 1946. Due to that, SMT was introduced in the 1970s, with improvement in accuracy of the translation. BI remains the most studied language (80%) for SMT as both the source or target language. Studies on MT or SMT for translating BM to other languages and vice versa are extremely rare. A few works on the BI-BM MT system in Malaysia have been around since the 1980s, however serious attention was not given to BM as the source language. To date, only three types of BM-BI translation engines are available. They are embedded in the Google translator, citcat.com, and the UTMK MT engine. These BM-BI MT engines still suffer poor quality output due to the ambiguity and complex grammar rules. We perceived the potential of semantic features to reduce such problems. This paper provides the basis for our claim that it is advisable to use semantic features in minimizing the problem of ambiguity and complex grammar rules, which in turn will improve the translation output quality.

For future work, an automated machine translation for BM-BI can be developed by embedding semantic properties as an effort to reduce ambiguity and complexity of grammar rules. In addition, PEnTrans (PEnT2) system can be further investigated since it has proven to have a better translation quality by scoring up to 85% accuracy. PEnTrans also enriched its grammar rules with semantic features.

REFERENCES

Ahmad Zaki, A. B., (1993). *Utilization of machine translation in Malaysia*. Retrieved April 2, 2010, from http://www.mt-archive.info/MTS-1993-Zaki.pdf

Ashizaki, T. (1995). *Adaptation of JICST's MT system for Workstation and PC's*. JICST (The Japan Information Center of Science and Technology). Retrieved January 16, 2010, from http://www.mt-archive.info/MTS-1995-Ashizaki.pdf

Bennett, W.S. & Slocum, J. (1985). *The LRC machine translation system*. Retrieved January 17, 2010, from http://www.mt-archive.info/CL-1985-Bennett.pdf

Chomsky, N. (1957). *Syntactic Structures*. Mouton: The Hague Paris.

Hutchins, J. (1986). *Machine Translation: Past, present, future.* Chapter 15: Artificial intelligent system. Retrieved December 20, 2009, from http://www.hutchinsweb.me.uk/PPF-15.pdf

Hutchins, J. (1995). *Machine Translation: a brief history*. Retrieved June 06, 2010, from http://aymara.org/biblio/mtranslation.pdf

Hutchins, J. (2005). *History of machine translation in a nutshell*. Retrieved December 20, 2009, from http://www.hutchinsweb.me.uk/Nutshell-2005.pdf

Hutchins, J. & Somers, H. L. (1992). *An introduction to machine translation: General introduction and brief history*. Retrieved January 18, 2010, from http://www.hutchinsweb.me.uk/IntroMT-1.pdf

Ikehara, S., Shirai, S., Yokoo, A., & Nakaiwa, H. (1991). *Toward an MT system without pre-editing-effects of new methods in ALT-J/E*. Retrieved January 16, 2010, from http://www.mt-archive.info/MTS-1991-Ikehara.pdf

Lau, P. (1987). *Eurotra: past, present and future*. Retrieved January 18, 2010, from http://www.mt-archive.info/Aslib-1987-Lau.pdf

Loh, S. & Kong, L. (1978). *An interactive on-line machine translation system (Chinese into English)*. Retrieved January 18, 2010, from http://www.mt-archive.info/Aslib-1978-Loh.pdf

Lee, Y., Yi, W., Seneff, S. & Weinstein, J. (2001). *Interlingua-based broad-coverage Korean-to-English translation in CCLINC*. Retrieved January 16, 2010, from http://www.mt-archive.info/HLT-2001-Lee.pdf

Lim, H. N., Ye, H. H., Lim, C. K. & Tang, E. K. (2007). Adapting an existing example-based machine translation (EBMT) system for new language pairs based on an optimized bilingual knowledge bank (BKB). Retrieved June 07, 2010, from

http://eprints.usm.my/9394/1/Adapting_an_Existing_Example-Based_Machine_Translation_(EBMT)_System_for_New_Language_Pairs_based_on_an_Optimized_Bilingual_Knowledge_Bank_(BKB).pdf

Muraki, K. (1987). *PIVOT: two-phase machine translation system*. Retrieved January 18, 2010, from http://www.mt-archive.info/MTS-1987-Muraki.pdf

Majlis Antarabangsa Bahasa Melayu (MABM). (2008). *Sejarah-latar belakang*. Retrieved February 02, 2010, from http://www.dbp.gov.my/lamandbp/main.php?Content=articles&ArticleID=136

Nagao, M., Tsujii, J., & Nakamura, J. (1985). *The Japanese government project for machine translation.* Retrieved January 17, 2010, from http://www.mt-archive.info/CL-1985-Nagao.pdf

Ornstein, J. (1955). *Mechanical Translation: new challenge to communication*. Retrieved January 18, 2010, from http://www.mt-archive.info/Ornstein-1955.pdf

O'Neill-Brown, P. (1996). *JICST Japanese-English machine translation system.* Retrieved January 16, 2010, from http://www.mt-archive.info/AMTA-1996-ONeill-Brown.pdf.

Sato, S. (1989). *Practical experience in the application of MT system*. Retrieved January 11, 2010, from http://www.mt-archive.info/MTS-1989-Sato.pdf

Saedi, C., Shamsfard, M., & Motazedi, Y. (2009). *Automatic translation between English and Persian texts*. Retrieved January 18, 2010, from http://www.mt-archive.info/MTS-2009-Saedi.pdf

Suhaimi, A. R, Noorhayati, A, Hafizullah, A. H, & Abdul Wahab, D. (2006). *Real Time on-line English-Malay machine translation (MT) system.* Retrieved on July 07, 2010, from http://www.cs.ieeemalaysia.org/RENTAS2006/papers/English-Malay-Translation-System.pdf

Tsujii, J. (1987). *The current stage of the Mu-Project.* Retrieved January 17, 2010, from http://www.mt-archive.info/MTS-1987-Tsujii-2.pdf

Yasuhara, H. (1993). *An example-based multilingual MT system in a conceptual language*. Retrieved January 17, 2010, from http://www.mt-archive.info/MTS-1993-Yasuhara.pdf