Proverb Treatment in Malay-English Machine Translation

Khirulnizam Abd. Rahman and Norita Mohd. Norwawi

Abstract -- Machine translation is a process by which computer software is used to translate a text from one natural language, in this case Malay, to another, English. They are normally short, generally known sentence of the folk which contains wisdom, truth, morals, and traditional views in a metaphorical, fixed and memorisable form and which is handed down from generation to generation. Although proverb do beautifies Malay literature, this brings challenges to machine translation since proverb cannot be translated literally, rather logically. This paper is to propose a research on proverb treatment in Malay-English machine translation. The treatment process started by identifying the Malay proverbs using the pattern matching approach. After the identification, the proverb will be translated into the meaning in English. To avoid literal (direct) translation a lookup table of the Malay proverb and the actual meaning in English has been prepared. There is another challenge, the ambiguity of the proverb meaning (there are Malay proverbs with several different meanings). To resolve this ambiguous issue, the possibility theory will be implemented with the help of the one of the WordNet feature, the semantic labeling.

Keywords-- idiom treatment , machine translation, Malay-English machine translation, pattern matching, possibility theory

I. INTRODUCTION

A. Multi Words Expressions (MWE)

MACHINE translation (MT) is sometimes defined as automated translation or machine aided translation. It is a process by which computer software is used to translate a text from one natural language, Malay, to another, English [27]. The study of Malay language in machine translation has been a serious topic since 1984 with the establishment of Unit Terjemahan Melalui Komputer (UTMK) in USM [5]. Malay language in the sense of this research is bahasa Melayu which is the standard first language in Malaysia. Although Indonesia, Malaysia and Brunei languages share the common roots of Austronesian family, the research is focusing into Malaysian Malay language.

Proverbs (peribahasa) in Malay language are beautiful elements to deliver advices, Malay teachings, moral values and comparison through metaphoric phrases [26]. They are normally short, generally known sentence of the folk which contains wisdom, truth, morals, and traditional views in a metaphorical, fixed and memorisable form and which is handed down from generation to generation. Although proverbs do beautifies Malay literature, however this brings challenges to machine translation since proverb cannot be translated literally, rather logically [7].

B. Brief Description of Malay Proverbs

Peribahasa or Malay proverb is a group of words in a fixed order that has a particular meaning that is different from the meanings of each word understood on its own [1]. There are four categories of Malay proverbs [1], which are simpulan bahasa, perumpamaan, bidalan and pepatah. Simpulan bahasa – normally consist of two words (sometimes three). The literal meaning of the word combination is different than the actual meaning of the ‘simpulan bahasa’. Example: Langkah kanan; literally means right footstep, yet the actual meaning is lucky.

Perumpamaan – phrases started with seolah-olah, ibarat, bak, seperti, macam, bagai or laksana.
Example: bagaikan pinang dibelah dua; literally means like betel nut split apart evenly, yet the actual meaning is compatible / equally beautiful and handsome for a pair of just married bride and bridgroom.

Pepatah – proverb that contains advices or teachings.
Example: Adat berperang, yang kalah jadi abu, menang jadi arang; literally means in war, loser become ashes, winner become coal, yet the actual meaning is in war, the defeated and the winner are both losers.

Bidalan – phrase (pepatah) started with jangan, biar or ingat.
Example: Kalau kail panjang sejengkal, lautan dalam jangan diduga; literally means if you have a short hook, do not attempt to fish in the deep sea, yet the actual meaning is if you have little knowledge, do not dare to dream big.

II. LITERATURE REVIEW

A. Multi Words Expressions (MWE)

MWE is combination of several words to form another meaning. As in other languages, Malay also contains a lot of MWE. Although some MWEs can be isolated in the tokenization process, and then analysed as a single cluster, most of them cannot [3].

Khirulnizam Abd. Rahman and Norita Mohd. Norwawi are with Universiti Sains Islam Malaysia Nilai, Negeri Sembilan, MALAYSIA email: khirulnizam@gmail.com
According to Tatabahasa Dewan [16], the basic sentence in Malay can be categorised into four basic structures (Table I).

<table>
<thead>
<tr>
<th>Malay Basic Sentence</th>
<th>Combination of Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noun Phrase + Noun Phrase</td>
</tr>
<tr>
<td>2</td>
<td>Noun Phrase + Verb Phrase</td>
</tr>
<tr>
<td>3</td>
<td>Noun Phrase + Adjective Phrase</td>
</tr>
<tr>
<td>4</td>
<td>Noun Phrase + Prepositional Phrase</td>
</tr>
</tbody>
</table>

Source: Tatabahasa Dewan Edisi Ketiga, 2008 [16]

Aiti Aw et al [2] in their research realized the important of Noun Phrases in Malay sentence structure decided to study this issue. They proposed a translation approach where they make use of parallel bilingual corpus to obtain a large set of bilingual terms and then use them to train a statistical engine. There’s another research by Rais et. al. [17] indexing the Malay MWE using combination of query translation approach and weighting schemes. The researchers did mention about dictionary is crucial in multiword detection.

B. Why proverbs are another type of MWE that are more difficult to translate?

Proverb and idioms are a part of MWE. The only problem with proverbs and idioms in MT is they cannot be translated literally, rather logically [7]. The translation machine needs to know the definite meaning of the proverbs.

C. Challenges in Malay proverbs automated detection.

These are several challenges that need to be consider in the automated detection of Malay proverbs.

a) Word with affixes – Example: “Kembang sayap” = “Mengembangkan sayap”.

b) Another word in between (stopword) Example: “berprijak di bumi nyata” or sometimes “berprijak di bumi yang nyata” which means “do not day-dreaming”.

c) Different proverb having the same meaning.

Example: “Baik budi”, “baik hati” which means kind-hearted. Another example: "bintang hati", "buah muka" – 1) face  2) pride

d) bagaici carik makan kapur - meaning 1) ashamed because of his own offense. 2) pleased.

e) Bagai cicak makan kapur - meaning 1) ashamed because of his own offense. 2) pleased.

D. Ambiguous meaning of Malay proverb.

There are Malay proverbs can have more than one meaning [18]. Examples of ambiguous Malay proverbs are:

a) “maita air” which means 1) lover, or 2) underground water resource.

b) “orang putih” which means 1) pious man, or 2) European people.

c) “air muka” – 1) face 2) pride

d) bawa diri - 1) running away  2) sulking  3) being independent.

e) Bagai cicak makan kapur - meaning 1) ashamed because of his own offense. 2) pleased.

f) ada air adalah ikan - meaning 1) there must be people in a country  2) fortune is everywhere

III. Problem Statement

The most challenging issue in interpreting natural language texts is the ambiguity problem [15], [10]. Proverbs are one part of the ambiguity issues. Proverbs normally come with fixed sequence of words; however the meaning is not based on the words directly [1]. Since proverb is translated logically, the machine translation algorithm needs to know the semantic (real meaning) of the phrases. On top of these issues, certain Malay proverbs have ambiguous meaning (more than one meaning) which the solution has not been mention in existing proverbs treatment [20], [7], [4].

IV. Research Questions

This study is to answer the following research questions;  
a) Is there any workable algorithm that can be implemented to detect Malay proverbs?
b) Since the nature of several Malay proverbs has ambiguity in the meaning, can this be solved using the existing algorithm?
c) Will the semantic analysis improve the Malay proverb translation?

V. The Research Objectives

The aim of this research is to propose a novel approach to proverb classification using Naïve Bayesian approach and translation into English using combination of dictionary mapping and semantic labeling.  
The objectives are:
a) To analyze the pattern of Malay proverbs.
b) To develop an algorithm to extract proverbs from Malay texts.
c) To develop an algorithm to translating the Malay idioms into English by providing the actual (logical) meaning (not literal meaning).
d) To model the semantic analysis in order to define ambiguous meaning of the Malay proverbs.
e) To evaluate the effectiveness of the proposed approach.

VI. The Scope of Research

The research is capable of detecting and defining the meaning of proverb in Malay text. Proverbs may appear in modern Malay literature such as in novel, short stories or blogs.

VII. The Research Contributions

The research’s major contributions are;
a) To develop an algorithm to identify proverb in Malay text.
b) To develop an algorithm that can translate Malay proverb by providing the right meaning of the proverbs that has ambiguous meaning.

VIII. Previous Works

These are several works by the previous researchers that have done remarkable jobs in solving the issue of proverbs/idioms detection and translation. These researches studied proverbs in many different languages such as Hindi, Punjabi, Malay, German, English and Italian.
Brahmaleen et. al. [4] suggested a method in identifying proverbs in Hindi text and translate them into Punjabi. The machine has a dictionary to map the Hindi proverb and the equivalent proverb in Punjabi. The illustration of the workflow of the said system is available in Figure 1.

Noah and Ismail [20] found out that Malay proverb can be detected using Naïve Bayesian model. The experiment suggested that the multinomial model shows slightly better performance as compared to the Multivariate Bernoulli model. This research suggested stop words removal to enhance the classification process.

There is another research from ATMA, UKM whereby they are providing a searchable database of Malay proverbs and idioms [25]. The application receives a complete or a part of proverb/idioms, using the SQL pattern matching to search the list of the proverbs/idioms, and the outputs are all the proverbs/idioms that similar to the user’s request.

Dmitra [7] studied the METIS II (in Figure 3) translation engine which capable of translating several languages such as Dutch, German, Greek and Spanish into English. The pre-translation process involves tokenization, lemmatization, tagging and chunker. By using the existing METIS II facilities, she managed to embed an idioms translation tool for Germany into English.

The idioms processor comprises of the following three components;
- a) German-English idioms dictionary,
- b) German source language corpus of sentences, and
- c) Syntactic matching rules.

The dictionary was constructed manually, the German corpus (containing idioms) assembled from varies prominent resources such as Europarl [9], sentence examples from the Web and DWDS [8]. The last component is the German syntactic matching rules which recognize the syntax of German idioms either in the form of continuous or discontinuous.

However, among all the researches mentioned in the discussion of the previous works, none of them mention about the treatment to the proverbs that have ambiguous meaning.

The research is an experimental based research whereby development of the system will be carried out, and evaluated to validate the performance. The following research processes are adopted:

Phase 1 : Literature Review (Information Gathering/ Acquisition)
A thorough literature review is conducted related to the area of research; covering theoretical, conceptual and developmental aspects. Current research on subject matter is analysed. The research gap is identified and refined to its potential venture.

Phase 2: Dictionary of Malay Proverb (and Semantic Labelling)
Dictionary of Malay Proverb is defined as the main dictionary used in the system. Suggested dictionary is the Peribahasa Melayu Kontemporari by Abdullah & Ainon, 2011 [1]. Semantic labelling (tagging) will be done manually to the list of proverbs that has ambiguous meaning (as suggested in Wordnet, 2012 [11] and Dmitra [7]).

Phase 3: Design
The design of system is carefully devised to support Malay proverb detection and translation. Pre-classification will involve tokenization, stop words removal and stemming. The proverb detection will be constructed from the pattern matching. As for the translation, direct dictionary lookup will be adopted (for proverb with single meaning). To resolve the ambiguity issue, the possibility theory will be implemented together with the semantic analysis and semantic labeling as in WordNet [11].

Phase 4: Development/ Implementation
The proposed system is developed based on the design phase aforementioned. Appropriate programming languages are used (suggestion to develop this system as a web-based application).

Phase 5: Evaluation
The system is tested in terms of its impact on the efficiency of the proverb classification, and correctness of translation. The testing dataset will be developed by combining sentences
that contains Malay proverb appear in Malay novel, short stories or blogs.

X. EXPECTED FINDINGS

Based on the problems identified, the researcher suggesting the following framework for Malay proverb treatment for Malay-English machine translation in Figure 3.

Tokenization is process of separating words and punctuation. This process provides a list of tokens which represents separated words and punctuations [7].

Stop words removal is a process of enlisting all the words that are considered not important (meaningless). In this case stop words are such as yang, itu, macam, seperti, bagai, laksana, itherat, umpama, ini, begitu, begini and etc. This process will produce better accuracy [22] in text classification.

Stemming is a process of finding the root of a word through removing the prefixes and postfixes. Word such as membawa is rooted from bawa which appear in membawa diri (or sometimes used as bawa diri). Stemming will provides more chances of finding the proverb in the dictionary.

Malay proverbs detection - by using the pattern matching approach, this process will identify Malay proverbs appearing in the text.

Proverb Semantic Analysis – in the case of proverb with ambiguous meaning, the semantic tagger will be analyzed.

XI. CONCLUSIONS

Proverb translation is challenging because the ambiguity and the logical definition. As highlighted in the previous discussion, there are Malay proverbs with ambiguous meaning. This research propose the following steps for proverb detection and translation of Malay text; tokenization, stemming, proverb detection and translation.

By developing the prototype of the system, the researcher could be able to minimize the confusion in defining proverb through automated translation. It is hoped that this approach will have great impact on the advancement of proverb treatment in Machine Translation system.

REFERENCES