An Approach to Handle Idioms and Phrasal Verbs in English to Myanmar Machine Translation System

Nyein Pyae Pyae Khin, Chaw Su Hlaing

nyeinpyaepyaekhin@gmail.com, 87chawsuhlaing@gmail.com
University of Computer Studies, Mandalay

Abstract

The structure of idioms and phrasal verbs are flexible in nature hence quite complex and challenging to handle in machine translation system. In this paper, an approach to handle idioms and phrasal verbs for English-Myanmar machine translation is proposed. This technique consists of two phases, the analyzing phase and the grouping phase. In the analyzing phase, the given English sentence is analyzed to find whether it contains any phrasal verbs or idioms. In the grouping phase, if any phrasal verb or idiom is found in the given sentence, it will be grouped into a single unit and it will be categorized with a special tag in order to denote it as the phrasal verb or idiom. The proposed approach can be incorporated with rule based English-Myanmar machine translation system with some modifications and can increase the accuracy of translation.

1. Introduction

Many resources such as news, stories and novels in English are being manually translated to Myanmar language. Of these, news and books from all around the world are translated from English to Myanmar language by human translators. Human translation is slow and also consumes more cost and time compared to machine translation. Hence, machine translation systems were developed to overcome the limitations of human translation. While translating from English to Myanmar, both idioms and phrasal verbs in English have more chances, to get translated to Myanmar in wrong sense. This is because of the idioms or phrasal verbs that convey individual meaning for each word in it instead of conveying a single meaning by considering it as a group of words while translating from English to Myanmar.

A phrasal verb creates a meaning different from its constituent verb [4]. It should not be translated by considering its constituent verb alone. Similarly an idiom conveys a peculiar meaning and cannot be predicted from the meaning of the constituent words. It

should be handled as a single unit during the translation process. The inconsistent machine translating for idioms and phrasal verbs have a great impact in the accuracy of English-Myanmar machine translation system. Since idioms and phrasal verbs are used more frequently in English language, it becomes necessary to handle idioms and phrasal verbs during the translation from English to Myanmar.

This paper is structured as follows: In the next section, the related works are reported. Nature of idiom and phrasal verb is presented in section 3. Section 4 describes the challenges in handling idioms and phrasal verbs. In section 5, the proposed technique is explained. The final conclusion is drawn in Section 6.

2. Related Work

This section discusses some related works of English-Myanmar machine translation system. Thinn Thinn Wai, Tin Myat Htwe and Ni Lar Thein proposed function tag and part-of-speech tag reordering rule extraction algorithms for English-Myanmar machine translation. In order to generate reordering rules; English-Myanmar parallel tagged aligned corpus is created. Additionally, reordering rules are generated automatically by using the linguistic information from this parallel tagged aligned corpus [1].

Handling of idioms and phrasal verbs has become one of the most important tasks to be handled in Machine Translation. Various approaches are developed to handle idioms and phrasal verbs in machine Translation ([2], [3] and [9]). Thiruumeni and Anand Kumar proposed an approach to handle idioms and phrasal verbs in English-Tami machine translation system in order to increase the accuracy of the translation. The BLEU and NIST scores calculated before and after handling the phrasal verbs and idioms during the translation process show a significant increase in the accuracy of the translation [2]. In this research, the comparisons of the translation of the input English sentence which contain idioms or phrasal verbs by the existing machine translation system with and without our proposed technique are described.

Martine Smets developed Machine Translation system to be able to handle verbal idioms. Verbal idioms become a challenge for machine translation systems because their meaning is not compositional, preventing a word-for-word translation, and they can be discontinuous, preventing a match during tokenization [3].

Iliana Simova and Valia Kordoni described an experimental evaluation of the significance of phrasal verb treatment for obtaining better quality statistical machine translation (SMT) results. Two ways of integrating phrasal verb information in a phrase-based SMT system are presented. Automatic and manual evaluations of the results reveal improvements in the translation quality in both experiments. Two integration strategies were used to incorporate information on the detected phrasal verb occurrences into a factored translation system. The first strategy encodes phrasal verbs as static units by concatenating their constituents via underscores. The second approach includes phrasal verb information into the translation table of the system in the form of a binary feature [9].

3. Nature of Idioms and Phrasal Verbs

Table 1. Types of Idioms with Examples

Types	Phrasal Verb	Meaning	Example
Separable	write down	Record	The professor instructed us to **write** the information **down**.
Inseparable	fill out	Complete	You need to **fill out** this form to register for the course.
Separable/ Inseparable	turn off	Stop	Please **turn** the TV **off**. (or) Please **turn off** the TV.

An idiom is a group of words with a special meaning which is different from the meanings of the separate words. In most cases when an idiom is translated, either its meaning is changed or it is meaningless. There are estimated to be at least 25,000 idiomatic expressions in the English language [6]. An idiom is generally a colloquial metaphor a term requiring some foundational knowledge, information, or experience, to use only within a culture, where conversational parties must possess common cultural references [7]. Therefore, idioms are not considered part of the language, but part of the culture. In linguistics, idioms are usually presumed to be figures of speech contradicting the principle of compositionality which states that the meaning of a complex expression is determined by the meanings of its constituent expressions. In general, idioms are based on pair of words, instrument, weather, color, etc. and are illustrated with examples in Table 1.

Phrasal verbs are verbs followed by an adverb or a preposition. In other words, it's the combination of the traditional verb and the adverb or preposition that creates a different meaning than if we were to examine them separately. Phrasal verbs can be broadly classified into two categories, transitive and intransitive. A transitive phrasal verb can either be followed by an object or it can contain an object between the verb and preposition or adverb and this can be further classified into separable and inseparable. Separable transitive phrasal verbs are those in which the object is placed between the verb and the preposition or adverb. Inseparable transitive phrasal verbs are those in which the object is placed after the preposition or adverb. Also there exist some transitive phrasal verbs that can be considered in both cases, separable and inseparable. Though some transitive phrasal verbs can be both separable and inseparable, the phrasal verb should take only the separable form when the object is a pronoun. Examples for the types of transitive phrasal verbs are illustrated in Table 2.

Table 2. Types of Transitive Phrasal Verbs

Type based on	Idiom	Meaning
Pair of words	Safe and sound	Undamaged
Instrument	As fit as a fiddle	In perfect health
Food	Full of beans	Full of life
Preposition	Down the drain	Wasted
Color	Out of the blue	Unexpectedly

An intransitive phrasal verb should neither be followed by an object nor should it contain an object between the verb and preposition or adverb. For example, He **passed away**.

4. Challenges in Handling Idioms and Phrasal Verbs

The main problem in machine translation system concerned phrasal verbs and idioms is that a phrasal verb is translated by considering the constituent verb in it, instead of considering it as a single unit. For example, the sentence "The manager *passed away*" will be translated as "မန်နေဂျာအဝေးသို့ကျော်ဖြတ်ခဲ့သည်" instead of "မန်နေဂျာကွယ်လွန်ခဲ့သည်". Here, the phrasal verb is translated in such a way that instead of conveying its meaning as a single unit i.e., '*to die*' conveys the meaning as '*to pass*' by considering the constituent verb in it. Similarly, an idiom is translated by considering the constituent words in it, without considering it as a single unit during the translation process from English to Myanmar. For example, the sentence "This work is *a piece of cake*" will be translated as "ဤအလုပ်သည်ကိတ်မုန့်တစ်တုံးဖြစ်သည်", instead of "ဤအလုပ်သည်အလွန်လွယ်ကူသည်". Here the idiom is translated in such a way that the translation conveys the literal meaning of constituent words in the idiom (i.e. '*a piece of cake*'), instead of conveying the meaning, '*easy*' by considering it as a single unit in the sentence. These examples above show how phrasal verbs and idioms affect the accuracy of the translation system.

In order to handle these idioms and phrasal verbs, a collection of most frequently used idioms and phrasal verbs have to be collected and manually translated to Myanmar. In such a way, idioms and phrasal verbs should convey the exact meaning or sense when considered it as a single unit in the sentence. Lexical dictionary for these idioms and phrasal verbs will be created with its equivalent translation in Myanmar. This dictionary can be referred by the machine translation system, if required, to replace the idioms or phrasal verbs in English with its equivalent Myanmar language.

5. Proposed Technique

The general block diagram of the proposed technique to handle the idioms and phrasal verbs in English-Myanmar machine translation system is shown in Figure 1. The input can be a sentence in English. The input sentence is firstly passed to the Phrasal verbs and Idioms Analyzer, before providing the input to the machine translation system. Here the input is thoroughly analyzed for any phrasal verbs or idioms in it, by looking up in the list of collected phrasal verbs and idioms.

If any phrasal verb or idiom is found in the sentence, it is passed to the Grouping phase. In the grouping phase, the words found in the analyzer phase are grouped together into a single unit and a special tag is assigned to it, so that this phrasal verb or idiom will be considered as a single unit during the whole translation process. In this grouping phase, when the transitive separable type of phrasal verb is grouped, the object in between the verb and the preposition or adverb is moved after the preposition or adverb in it. For example, the sentence, "Please **turn** the light **on**." will be grouped as "Please **turn on** the light." and will be translated as "ကျေးဇူးပြု၍ မီးဖွင့်ပေးပါ". In this way, the phrasal verbs are handled similar to the traditional verbs.
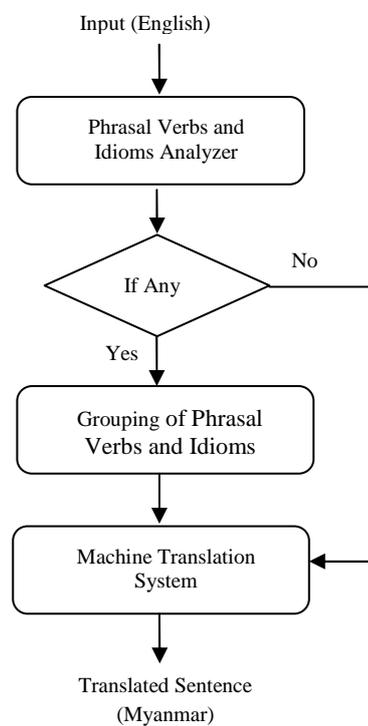

Figure 1. General Block Diagram for the Proposed Technique to Handle Phrasal Verbs and Idioms

The above block diagram for the proposed technique can be integrated to any English-Myanmar rule based machine translation system with some modifications. The following section will give a clear idea of how this technique can be used in rule based machine translation system.

5.1 Rule Based Machine Translation System with Proposed Technique

In rule based machine translation system, the given English sentence annotated with lemma, part of

speech tag, morphological and dependency information is passed to the Phrasal verbs and Idioms Analyzer phase before passing it to the actual translation process. In this phase, the analyzer checks the given sentence whether it contains any idioms or phrasal verbs or not. If found, the sentence is passed to the grouping phase where the words found in the previous phase are grouped together as a single unit in the sentence and it is assigned with a special tag "PHV" for phrasal verbs and "IDM" for idioms along with the annotated part of speech tag information.

Input (English)

```
        Phrasal Verbs and
        Idioms Analyzer
```

```
           If Any              No
```

Yes

```
        Grouping of Phrasal
        Verbs and Idioms
```

```
  English-
  Myanmar
  Phrasal Verbs        Rule Based Machine
  and Idioms           Translation System
  Lexical
  Dictionary
```

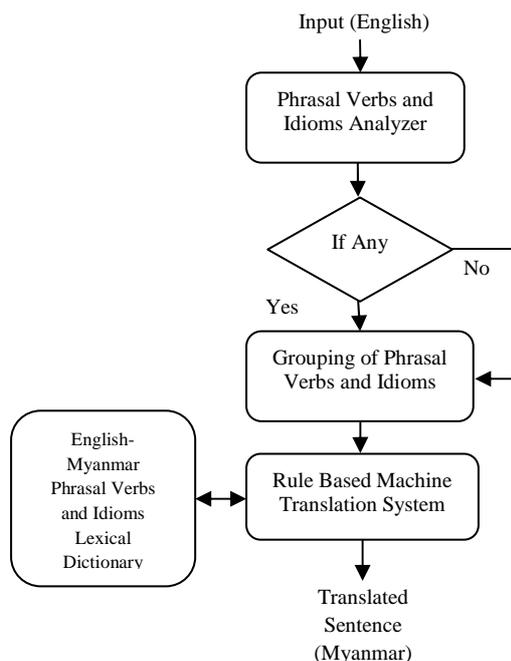Translated
Sentence
(Myanmar)

Figure 2. Block Diagram for the Proposed Technique to Handle Phrasal Verbs and Idioms in Rule Based English-Myanmar Machine Translation System

In the case of phrasal verbs, transitive and intransitive, the structure is differentiated by the object following it or between the verb and adverb or preposition. If it is intransitive, an asterisk symbol is added to the end of the original phrasal verb. So that the two forms of the phrasal verb can be differentiated easily while translating. For example, intransitive form of the phrasal verb "*drop by*" will be changed to "*drop-by**" which means "*visit without an appointment*".

During the translation process, the unit assigned with the special tag "PHV" will be handled as verb indeed, but during lexical replacement of English to Myanmar, instead of retrieving from the lexical dictionary for verb, some modification has to be made in the existing system so that it retrieves from lexical

dictionary for phrasal verbs and for the words with the tag "IDM", the lexical replacement has to be made from the lexical dictionary for idioms. The block diagram for the technique for English-Myanmar rule based machine translation system is shown in Figure 2.

6. Conclusion

In this paper, an approach to handle idioms and phrasal verbs in rule based English-Myanmar Machine Translation System is proposed. In order to produce more consistent Myanmar sentence, the input English sentence will be passed the two phases, Phrasal Verbs and Idioms Analyzer phase and Grouping phase. This research attempts to provide a significant increase in the accuracy of English-Myanmar machine translation with a set of idioms and phrasal verbs. The findings of this research will greatly benefit further NLP steps of English-Myanmar machine translation system.

References

[1] Thinn Thinn Wai, Tin Myat Htwe and Ni Lar Thein, "Automatic Reordering Rule Generation for English-Myanmar Translation", University Press, 2011, Yangon, Myanmar.

[2] Thiruumeni P G, Anand Kumar M, Dhanalakshmi V and Soman KP, "An Approach to Handle Idioms and Phrasal Verbs in English-Tami Machine Translation System" International Journal of Computer Applications (0975 – 8887) Volume 26, No.10, July 2011

[3] Breidt, E., Segond F and Valetto G. 1996. "Local grammars for the description of multi-word lexemes and their automatic recognition in texts", Proceedings of COMPLEX96, Budapest.

[4] Courtney, Rosemary. 1989. "Longman Dictionary of Phrasal Verbs". Longman Group UK Limited, ISBN 0-582- 55530-2 CSD, ISBN 0-582-05864-3 PPR.

[5] The Oxford Companion to the English Language. 1992. pp.495–96.

[6] Jackendoff, R. 1997. "The architecture of the language faculty". Cambridge, MA: MIT Press.

[7] Chunli Yang 2010. "Cultural Differences on Chinese and English Idioms of Diet and the Translation", In CCSENET, English Language Teaching.

[8] Thinn Thinn and Ni Lar Thein, "Markov-based Reordering Model for English-Myanmar Translation", University press, 2011, Yangon, Myanmar.

[9] Iliana Simova and Valia Kordoni "Improving English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment", Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.