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

Thiruumeni P G, Anand Kumar M
Computational Engineering & Networking,
Amrita Vishwa Vidyapeetham,
Coimbatore, TN, India

Dhanalakshmi V, Soman K P
Computational Engineering & Networking,
Amrita Vishwa Vidyapeetham,
Coimbatore, TN, India

ABSTRACT

In this paper, we report our work on incorporating a technique to handle phrasal verbs and idioms for English to Tamil machine translation. While translating from English to Tamil, both phrasal verbs and idioms in English have more chances, to get translated to Tamil 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 Tamil. This in turn affects the accuracy of the translation. The proposed technique is used to handle the idioms and phrasal verbs during the translation process and it increases 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. This technique, proposed for English to Tamil machine translation system, can be incorporated with machine translation system for English to any language.

General Terms

Natural Language Processing, Machine Translation.

Keywords

Phrasal verbs, Idioms, Statistics Machine Translation, Rule based Machine Translation.

1. INTRODUCTION

Machine translation is an important and most appropriate technology for localization in a linguistically diverged country like India [1]. The reason for choosing automatic machine translation rather than human translation is that machine translation is better, faster and cheaper than human translation. Many resources such as news, weather reports, books, etc., in English are being manually translated to Indian languages. Of these, News and weather reports from all around the world are translated from English to Indian languages by human translators more often. Human translation is slow and also consumes more time and cost compared to machine translation. Hence, there is a good scope for machine translation to overcome the human translation, in near future. There are machine translation systems that are being developed in order to translate from English to Indian languages. But there are problems that make these systems not able to produce a good translation of text from English to Indian languages. Here we incorporate the technique with English-Tamil machine translation system.

One of the problems in English-Tamil machine translation system is to handle the idioms and phrasal verbs. A phrasal verb, which is a combination of a verb and a preposition or adverb, creates a meaning different from its constituent verb. It should not be translated by considering its constituent verb alone. Similarly an idiom, which is usually a group of words, 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. But the existing machine translation system handles the translation of a phrasal verb by translating the constituent verb in it and idiom by translating each constituent word in it. This makes idioms and phrasal verbs to have a great impact in the accuracy of English-Tamil machine translation system.

We describe a technique that can be used to handle idioms and phrasal verbs which can increase the accuracy of English - Tamil translation, when incorporated with any existing English - Tamil machine translation system. The technique consists of two phases, analyzing phase and grouping phase. In analyzing phase, the given English sentence is analyzed to find whether it contains any phrasal verbs or idioms. In grouping phase, if the given sentence is found to contain a phrasal verb or an idiom, then 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. This tag will be considered instead of the part-of-speech tag during the translation process. This approach can be used in both rule based and factored statistical machine translation with some modifications.

2. RELATED WORKS

Handling of idioms and phrasal verbs is one of the most important tasks to be handled in Machine Translation. Various approaches are developed to handle idioms and phrases in machine Translation. Sahar Ahmadi and Saeed Ketabi focused on analyzing the translatability of color idiomatic expressions in English- Persian and Persian- English texts to explore the applied translation strategies in translation of color idiomatic expressions and also to find cultural similarities and differences between color idiomatic expressions in English and Persian [2]. Martine Smets et al. developed their Machine Translation system in such a way such that it handles the verbal idioms. Verbal idioms constitute a challenge for machine translation systems: their meaning is not compositional, preventing a word-for-word translation, and they can be discontinuous, preventing a match during tokenization [3].

Table 1. Types of Phrasal Verbs with Examples

Type		Phrasal Verb	Meaning	Example
Transitive	Separable	cut * off	interrupt someone while they were speaking	She cut him off while he was talking.
	Inseparable	look into +	Investigate	The police are looking into the murder.
	Separable / Inseparable	pass * out +	Distribute	We need to pass these sweets out. *(Separable)* We need to pass out these sweets. *(Inseparable)*
Intransitive		pass away	Die	He passed away.

** - Object in between, + - Object after the verb and preposition or adverb*

Elisabeth Breidt et al. suggested describing their syntactic restrictions and their idiosyncratic peculiarities with local grammar rules, which at the same time permit to express regularities valid for a whole class of multi-word lexemes such as word order variation in German [4]. Digital Sonata provides NLP services and products. It has released its tool kit called, Caraboa Language Kit, in which idioms serves as the backbone of its architecture and it is mainly rule based. Here, the idioms are considered as sequences and each sequence is a combination of one or more lexical units.

3. PHRASAL VERBS AND IDIOMS – AN OVERVIEW

As described earlier, a phrasal verb is a combination of a verb and a preposition or adverb that creates a meaning different from its original constituent verb [5]. 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. An intransitive phrasal verb should neither be followed by an object nor should it contain an object between the verb and preposition or adverb. Examples for the types of phrasal verbs are illustrated in Table 1.

An idiom is usually a group of words whose meaning will be peculiar and cannot be predicted from the meanings of the constituent words [6]. Also, it can be considered as an expression that is not readily analyzable from its grammatical construction or from the meaning of its component parts. In other words, an idiom is an expression, word, or phrase whose sense means something different from what the words literally imply [7]. 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 [7] in the English language. An idiom is generally a colloquial metaphor [8] a term requiring some foundational knowledge, information, or experience, to use only within a culture, where conversational parties must possess common cultural references. 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, number, nationality, color, etc. and are illustrated with examples in Table 2.

Table 2. Types of Idioms with Examples

Type based on	Example	Meaning
Pair of words	Safe and sound	Undamaged, safe
Numbers	To have second thoughts	To form an opinion after reconsidering
Names	Jack of all trades	Person who has an ability to do a lot of different jobs
Food	A piece of cake	Very easy
Color	Out of the blue	Unexpectedly
Prepositions	Out of date	No longer in use or fashion

4. CHALLENGES IN HANDLING IDIOMS AND PHRASAL VERBS

The main problem in existing machine translation system due to 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 minister *passed away*” will be translated as “amaiccar *thUram thErcciyatainthAr*” (அமைச்சர் தூரம் தேர்ச்சியடைந்தார்) instead of “amaiccar *iyaRkai eythinAr*” (அமைச்சர் இயற்கை எய்தினார்). 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 and an idiom is translated by considering the constituent words in it, instead of considering it as a single unit during the translation process from English to Tamil. For example, the sentence “This work is *a piece of cake*” will be translated as "intha vElai inirottiyin oru pakuthiyAkum" (இந்த வேலை இனிரொட்டியின் ஒரு பகுதியாகும்), instead of "intha vElai *eLithAnathu*" (இந்த வேலை எளிதானது). 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. As idioms cannot be analyzed from its grammatical construction, handling the idioms in translation process becomes a challenging task. Since idioms and phrases are used more frequently in English language, it becomes necessary to handle the idioms during the translation from English to Tamil.

In order to handle these phrasal verbs and idioms, a collection of most frequently used phrasal verbs and idioms have to be collected and manually translated to Tamil in such a way that it should convey the exact meaning or sense of the phrasal verb or idiom when considered as a single unit in the sentence. Lexical dictionary for these phrasal verbs and idioms is created with the collected phrasal verbs and idioms and its equivalent translation in Tamil. This dictionary can be referred by the machine translation system, if required, to replace the phrasal verbs or idioms in English with its Tamil equivalent. While creating the lexical dictionary for phrasal verbs, the dictionary is created with root form of the phrasal verbs, so that all the inflections of the phrasal verbs can be handled in a way similar to that of verbs. For example, instead of '*passed away*' its root form '*pass away*' is added to the lexical dictionary.

Also in order to handle the separable transitive phrasal verbs, some rules have to be coded such that in case of phrasal verbs which can be both separable and inseparable and if it have pronoun as the object, it should be handled as separable. Some of the phrasal verbs convey one meaning when they are transitive, which is entirely different from the meaning when they take intransitive form. For example, the phrasal verb '*show up*' gives the meaning 'make someone seem inferior' in transitive case and 'arrive without prior notice' in intransitive case. These cases are handled by taking the object in consideration, so that it distinguishes the transitive and the intransitive form of the phrasal verb during the translation process.

5. IMPLEMENTATION

The general block diagram of proposed technique to handle the phrasal verbs and idioms during English-Tamil machine translation system is shown in Fig 1. The input to this technique can be a sentence in case of rule based machine translation and bilingual and monolingual corpus for training and input sentences in case of statistical machine translation. Before providing the input to the machine translation system for further process, the input is passed to the first phase of the proposed technique, Phrasal verbs and Idioms Analyzer. Here the input is thoroughly analyzed for any phrasal verbs or idioms in it, by looking up in the list of phrasal verbs and idioms collected. If any phrasal verb or idiom is found to be in the sentence then it is passed to the second phase of the technique, the grouping phase. In the grouping phase, the words in the phrasal verb or idiom that is found to be in the input 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.

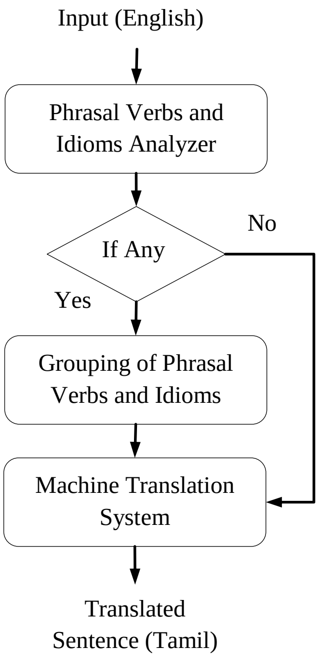
Input (English)
Phrasal Verbs and Idioms Analyzer
If Any
No
Yes
Grouping of Phrasal Verbs and Idioms
Machine Translation System
Translated Sentence (Tamil)

Fig 1: General block diagram for the Proposed Technique to handle Phrasal Verbs and Idioms

In the grouping phase, while grouping the words in the phrasal verb which is of transitive separable type, the object in between the verb and the preposition or adverb is moved after the preposition or adverb in it. For example, the sentence, "She *cut him off* while he was talking" will be grouped as "She *cut-off him* while he was talking" and will be translated as "avan pEcikkoNtirukkum pozuthu avaL avanai kURukkittaL" (அவன் பேசிக்கொண்டிருக்கும் பொழுது அவள் அவனைக் குறுக்கிட்டாள்), as the phrasal verbs are handled in the way similar to verbs. Lexical dictionary with 900 idioms and 241 phrasal verbs have been created for idioms and phrasal verbs, separately. The above block diagram for the proposed technique can be integrated to any English-Tamil rule based machine translation system or to any English-Tamil statistical machine translation, with some modifications in the general technique. The following section will give a clear idea of how this technique can be used in rule based and factored statistical machine translation.

5.1 Rule Based Machine Translation System

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 first-phase of the technique, Phrasal verbs and Idioms analyzer phase, before passing the sentence to the actual translation process. In this phase, the analyzer checks for any phrasal verbs or idioms present in the given sentence. If found, the sentence is passed to the grouping phase, where the words that form the phrasal verb or idiom found in the analyzer 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.

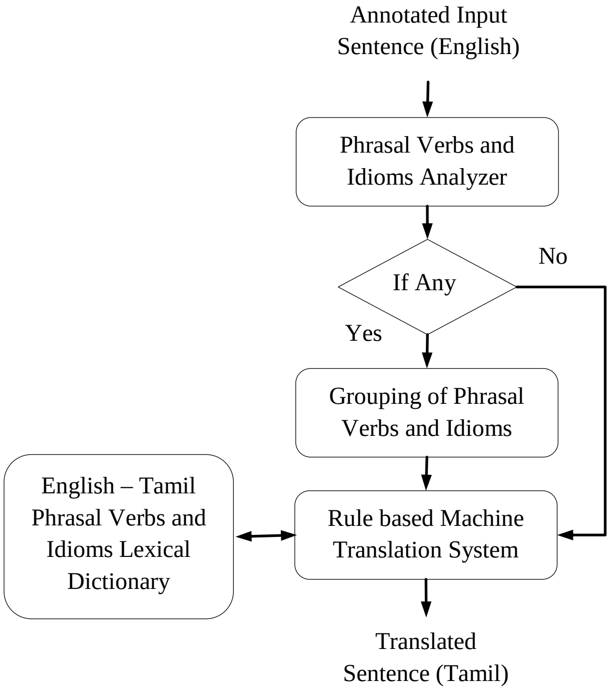

Fig 2: Modified block diagram for the Proposed Technique to handle Phrasal Verbs and Idioms in Rule Based English-Tamil Machine Translation.

In case of phrasal verbs which take both transitive and intransitive form, the form of the phrasal verb is differentiated by the object following it or in between the verb and adverb or preposition. An asterisk symbol is added to the end of root of the phrasal verb, if it is intransitive. So that while translating, the two forms of the phrasal verb can be differentiated easily. For example, intransitive form of the phrasal verb '*show up*' will be changed to '*show-up**' which means '*arrive without prior notice*'. All other annotated information of the words grouped to form a single unit is also grouped in the sequence of the words as in the phrasal verb or idiom. 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 Tamil, 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 modified technique for English-Tamil rule based machine translation system is shown in Figure 2.

5.2 Factored Statistical Machine Translation System

In the existing factored statistical machine translation system [9] before the training phase the bilingual and monolingual corpus is pre-processed by the proposed technique to group the phrasal verbs and idioms in to a single unit. Here, the term factored means the corpus along with information such as lemma, part-of-speech tag and morphological information for each word in every sentence in the corpus. The statistical machine translation decoder translates the sentences from English to Tamil by considering the factored information as translation factors [10]. Here, the technique has been modified so that in the proposed technique's analyzer phase, the English sentences are analyzed for phrasal verbs or idioms. If found, in the grouping phase the phrasal verbs or idioms in English as well as its equivalent in Tamil are also grouped into a single unit. Also the Tamil monolingual corpus has been analyzed for phrasal verbs or idioms, and grouped into a single unit, if found any. And the part-of-speech category for phrasal verbs and idioms are assigned as 'PHV' and 'IDM' respectively. The technique is applied in a similar way to the monolingual corpus. After the grouping phase of the technique, the bilingual and monolingual corpus is passed to the training phase of the decoder. During the testing phase, the factored sentence is pre-processed by this technique and then passed to the decoder for translation. The output of the decoder is given to the morphological generator to generate the final translated sentence. The block diagram for the modified technique for English-Tamil factored statistical machine translation system is shown in Figure 3.

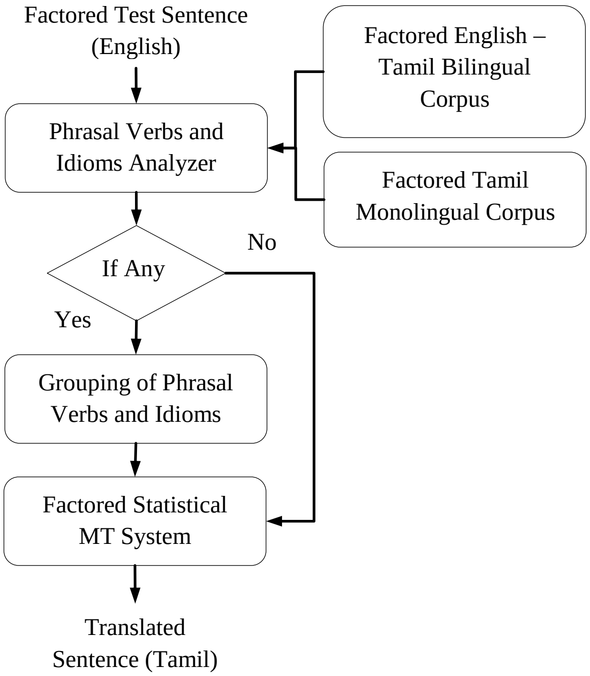

Fig 3: Modified block diagram for the Proposed Technique to handle Phrasal Verbs and Idioms in Factored English-Tamil Statistical Machine Translation.

6. RESULTS AND CONCLUSION

The machine translation system for English-Tamil has been tested and evaluated for four cases, (1) the baseline machine translation system, (2) the baseline machine translation system with the proposed technique to handle phrasal verbs, (3) the baseline machine translation system with technique to handle idioms and (4) the baseline machine translation system with technique to handle both phrasal verbs and idioms, in both the rule based and factored statistical machine translation system.

Table 3. Evaluation Results

	Machine Translation System	BLEU	NIST
Rule-Based	Baseline (BL)	40.09	7.11
	BL + Phrasal Verbs(PHV)	49.13	8.46
	BL+ Idioms (IDM)	46.01	7.89
	BL+ PHV+ IDM	54.26	9.05
Factored SMT	Baseline (BL)	47.62	8.29
	BL + Phrasal Verbs(PHV)	53.81	9.37
	BL+ Idioms (IDM)	51.66	9.00
	BL+ PHV+ IDM	58.94	9.46

The rule based machine translation system has been evaluated with a test data set of 500 sentences. The factored statistical machine translation system has been trained with English – Tamil bilingual corpus with 20,000 parallel sentences and a Tamil monolingual corpus of 50,000 sentences and has been evaluated with another test data set of 500 sentences. Both the systems have been evaluated for the four cases with BLEU and NIST score and the results have been tabulated in Table 3, which shows that incorporating this technique to handle idioms and phrasal verbs has increased the accuracy of the existing English - Tamil machine translation systems. Comparison of how the sentences containing phrasal verbs or idioms in English gets translated to Tamil with the existing machine translation system and the existing machine translation system with the proposed technique to handle the phrasal verbs and idioms are illustrated with examples in Table 4.

7. FUTURE WORK

The current proposed technique provides a significant increase in the accuracy of translation with a small set of idioms and phrasal verbs. Hence more number of phrasal verbs and idioms has to be added to the phrasal verbs and idioms lexical dictionary in order to improve the accuracy of the translation, further. We propose to incorporate this technique with English to other Indian language machine translation system and evaluate how this technique helps in increasing the accuracy of the machine translation systems.

8. ACKNOWLEDGEMENTS

This work was part of the '*Machine Translation for English to Tamil Language*' project funded by CDAC, Pune and Department of Information Technology (DIT), India. We would also like to thank them for their support to complete this work successfully.

Table 4. Comparison of translation results of Machine Translation System with and without the proposed technique to handle phrasal verbs and idioms

Phrasal verbs or Idioms		English	Output of Baseline System	Output of Baseline System with proposed technique
Phrasal Verbs	Account for	He should account for his mistakes	அவன் அவனுடைய தவறுகளுக்கு எண்ண வேண்டும்	அவன் அவனுடைய தவறுகளுக்கு விளக்கமளிக்க வேண்டும்
	Call off	The meeting was called off	கூட்டம் அழைக்கப்பட்டது	கூட்டம் ரத்தானது
	Pass out	He passed the sweets out	அவன் தேர்ச்சியடை இனிப்பான	அவன் இனிப்புகளை வினியோகித்தான்
Idioms	Jack of all trades.	Arun is a jack of all trades	அருண் அணைத்து வர்த்தகங்களுக்கும் ஒரு ஜேக்	அருண் ஒரு சகலகலா வல்லவன்
	A piece of cake	This job is a piece of cake	இந்த வேலை இனிரொட்டியின் ஒரு பகுதியாகும்	இந்த வேலை எளிதானது
	Smell a rat	I smell a rat on seeing him	நான் அவனை கண்டவுடன் ஒரு எலியை நுகர்ந்தேன்	நான் அவனை கண்டவுடன் சந்தேகமடைந்தேன்

9. REFERENCES

[1] Durgesh Rao, 2001. “Machine Translation in India: A Brief Survey”. In Proceedings of the *SCALLA 2001 Conference*, Bangalore, India.

[2] Sahar Ahmadi and Saeed Ketabi. 2011. “Translation Procedures and problems of Color Idiomatic Expressions in English and Persian”. The Journal of *International Social Research*, Volume: 4 Issue: 17.

[3] Martine Smets, Joseph Pentheroudakis and Arul Menezes. “Translation of verbal idioms”, *Microsoft Research*.

[4] 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.

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

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

[7] Jackendoff, R. 1997. “The architecture of the language faculty”. Cambridge, MA: MIT Press.

[8] Chunli Yang 2010. “Cultural Differences on Chinese and English Idioms of Diet and the Translation”, In *CCSENET*, English Language Teaching.

[9] Anand Kumar M, Dhanalakshmi, Soman K P and Rajendran S. 2011. “Morphology based Factored Statistical Machine Translation System for English to Tamil”, *10th Tamil Internet Conference, INFIT & UPEN*, Philadelphia, USA.

[10] Philipp Koehn and Hieu Hoang. 2007. “Factored Translation Models”. In Proceedings of the *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 868–876, Prague.