International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Development of English-Punjabi Parallel Corpus for Idioms and Phrases using Automatic Text Alignment Technique

Jitin Chhabra[*1], Dharam Veer Sharma[2]
[*1]M.Tech (CSE), Department of Computer Science, Punjabi University, Patiala, India
[2]Associate Professor, Department of Computer Science, Punjabi University, Patiala, India

ABSTRACT

Machine Translation has gained a widespread attention in the area of Natural Language Processing due to the increasing Human-Computer interactions in the recent past. Machine Translation provides assistance in the communication among different cultural languages using its rule based and statistical methods of linguistic computations. The Machine Translation systems made so far faces some hurdles in the understanding and processing of multi word fixed expressions like Idioms and Phrases. In order to provide the better efficiency in the translation process of Idioms and Phrases, we proposed the development of Parallel Corpus using Automatic Alignment Technique. We implemented the Alignment Technique on the Idioms and Phrases of English and Punjabi Scripts. We have used the Adjectives from the Tokenized Words of expressions as an element of context identifier and a bilingual English-Punjabi dictionary for the translation process. We have performed the Alignment Experiment and found the mappings for the development of parallel corpus. The resulted set of mappings has been matched with the original equivalents and the wrong sets of mappings are filtered out. The results also pointed out the problems of context identification and ambiguity of words which give rise to the wrong set of mappings.

Keywords : Machine Translation, Parallel Corpus, Idioms and Phrases, Word level matching, Context Identification, Automatic Text Alignment.

I. INTRODUCTION

With the widespread use of computer systems and their interactions with the humans, Machine Translation has gained a massive attention in the area of natural language processing in the recent past. Machine Translation is assistance for communication among different cultural languages. From its variants including rule based approach and statistical approach, the latter has gained much importance as it gives more sound results. The use of parallel text corpora and automatic text alignment models are the vital ingredients in the statistical approach of machine translation which includes the correspondence between the words in source language and the target language.

II. NATURAL LANGUAGE PROCESSING

Natural Language Processing is a field of artificial intelligence, computational linguistics and computer science, related with the interactions that occur between computer systems and languages known by humans. Natural language processing includes the study of computational as well as mathematical modeling of various parts of a situation in a language and the development of a broad range of computational systems. Natural language processing is having a vital function in the field of computer science as many views of this area work with linguistic characterization of computation. This contains the spoken language computational systems that integrate speech and human languages. Natural language processing is the region of research and development studies which examines how computers could be used to manipulate and understand human language text/speech to perform useful things. Natural language processing applications include regions of study, such as machine translation systems, natural language text/speech processing and recognition, interfaces for user terminals, information

retrieval systems for cross language platforms, artificial intelligence and expert knowledge systems.

III. MACHINE TRANSLATION

Machine Translation is an area of computational linguistics that illustrates is an act of using software programs to translate text and speech from one human language to another human language. Machine Translation carries out a simple task of translating words, sentences and phrases from source language to the target language. Machine Translation describes the skill of computer systems to perform the job of translation between different human known natural languages. The construction of multilingual Machine Translation systems for any human languages using electronic resources and tools is a quite difficult job that requires a lot of computational knowledge and skills.

Machine translation can be performed with rule-based, statistical-based or hybrid methods. Machine Translation systems are specially developed for two particular languages, called a bilingual translation system, and for more than a single pair of languages, known as multilingual translation systems. A bilingual translation system can be unidirectional which converts text from one human Language to another human Language, or it can be bidirectional capable of inter-lingual translations. The approaches differ in the examination of the source language and degree of realization to reach a language independent representation of context between the source and target human languages. Obstacles in efficient and precise Machine Translation output can be characterized to ambiguities in the natural languages.

IV. PARALLEL CORPUS

A parallel corpus is basically a collection of texts in different languages where one of them is the original text and the other is their translations. Corpora in general and, particularly, parallel corpora are vital resources for tasks in the translation field like linguistic studies, information retrieval systems development or natural language processing. In order to be useful, these resources must be available in reasonable quantities, because statistics are used in most application methods. The quality of the findings depends a lot on the size of the corpora, which means robust tools are needed to build and process them. The alignment at sentence and word levels makes parallel corpora both more interesting and more useful. As long as parallel corpora exist, sentence aligned parallel corpora is an issue which is solved by sentence aligners. Some of these tools are available as open-source software, while others have free licenses for non-commercial use, and produce reasonable results.

V. PARALLEL CORPORA ALIGNMENT

The first phase in fetching useful information from bitexts is to find corresponding words and/or text segment boundaries in their two halves called bitext Maps. Without an automatic method for matching corresponding text units in their two halves bilingual texts are of no use. Although we can add morphological analysis, word lemmas, syntactic analysis and so on to parallel corpora, these properties are not specific to parallel corpora. The first step to enrich parallel corpora is to enhance the parallelism between units on both texts. This process is called "alignment". Alignment can be done at different levels, from paragraphs, sentences, segments, words and characters. Usually, alignment tools perform the alignment at sentence and word levels:

- Texts are sequences of sentences. To sentence align two texts is to create relationships between related sentences.
- The same idea can be used for the word alignment process as well. Sentences are sequences of words. So, the word alignment process will add links between words from the original and the translated text. Word alignment can be viewed in two different ways:–

1) For each word, in a sentence, find the corresponding word in the translated sentence. This means that, for each occurrence of a word, it has a specific word linked to it.

2) For each word from the source corpus, find a set of possible translations (and its probability) into the target corpus. This leads to a Probabilistic Translation Dictionary (PTD), where for each different word of the corpus we have a set of possible translations and their respective probability of correctness.

VI. IDIOMS AND PHRASES

The common sentence 'Idioms and Phrases' refers to commonly used unit of words in English. Idioms are

used for informal situations, whereas phrases may also be rather applied in formal. Idioms and Phrases play a crucial role in learning any language. Idioms and phrases are the multiword fixed expressions which are used in an idiomatic sense, rather than a figurative meaning under specific scenarios. Idioms are often regarded full sentences whereas Phrases are made up of a few words which are used as a grammatical component in a sentence.

VII. HURDLES FACED DURING THE TRANSLATION OF IDIOMS AND PHRASES

As the two languages are quite different in the cultural aspects, grammatical word orders, word sense ambiguities, their translation process needs to be taken care of these considerations. The challenges faced by the designer of an algorithm for automatic text alignment during the machine translation of multiword fixed expressions are as follows:

a) A fixed expression or an idiom may be having no equivalent in the target language.

Example: The following idioms does not have English equivalent idioms:-

ਜਾਮ ਦੀ ਰੰਨ ਤੇ ਬਿੱਲੀ ਦੇ ਕੰਨ |

Idiomatic meaning- Wealth without use is no wealth.

ਅੰਨ੍ਹੇ ਕੁੱਤੇ,ਹਿਰਨਾਂ ਦੇ ਸ਼ਿਕਾਰੀ |

Idiomatic meaning- Incompetent persons cannot do a tough job.

ਸ਼ੇਰਨੀ ਦਾ ਦੁੱਧ ਸੋਨੇ ਦੇ ਭਾਂਡੇ 'ਚ ਹੀ ਸਮਾਉਂਦਾ ਐ |

Idiomatic meaning- Extraordinary care is needed for unique things.

ਆਲਸੀਆਂ ਦੇ ਪਿੰਡ ਵਿਹੜੇ ਨੀ ਹੁੰਦੇ |

Idiomatic meaning- Lazy people lives in the same society.

b) A fixed expression or an idiom might be having a similar expression in the target language, which resembles the source text but its context of use may be different.

Example:

English: To kick the bucket. (To die)

Punjabi: ਲੱਤ ਮਾਰਨੀ | (ਵਿਘਨ ਪਾਉਣਾ)

The above example shows how both languages have similar looking expressions but their intentions and context are different.

c) An idiom may be used in the source text in both its idiomatic and literal sense at the same time.

Example: The following idioms have literal as well as figurative meaning.

1) Kick the bucket.

This idiom has figurative meaning as to die and literal meaning as to really kick the bucket.

2) Spill the beans.

This idiom has figurative meaning as to tell/leak secret information and literal meaning as to really spill them.

d) The rule of building idioms in written form, the contexts of using it, and the frequency in which it is used may be different in the source and target languages.

Example:

English: It's raining like cats and dogs.

Punjabi: ਟੋਆ-ਟਿੱਬਾ ਇਕ ਹੋਣਾ |

English: A drop in the ocean.

Punjabi: ਊਠ ਦੇ ਮੂੰਹ ਜੀਰਾ ਦੇਣਾ |

e) Recognition and interpretation problems.

Example: Break a leg.

This idiom describes the meaning as to wish good luck and hence difficult to interpret as it does not have anything linked with breaking the leg.

f) Ambiguity problems.

Example:

ਅੱਖਾਂ ਵਿਖਾਉਣਾ: ਡਰਾਉਣਾ |

English Meaning: To make someone feel afraid or frightened.

ਅੱਖਾਂ ਵਿਖਾਉਣਾ: ਡਾਕਟਰ ਤੋਂ ਅੱਖਾਂ ਦਾ ਮੁਆਇਨਾ ਕਰਾਉਣਾ |

English Meaning: Eye checkup from doctor.

VIII. PROBLEM FORMULATION

The multiword fixed expressions like Idioms and phrases are non isolating components, found in abundance in most of the languages in the world. As the semantics and meaning of these arrangement of words cannot be explained from the individual meanings of the component words constituting them, so there are some obstacles in the process of understanding as well as machine translation. The action of translating idioms and phrases from one human language into its counterpart is a highly intelligent task which demands a translator to have a sound knowledge and understanding of both the languages and cultures being communicated in the process of aligning and finding a precise equivalent for the inter-lingual idiomatic pairs. The Machine Translation systems made so far has not coped up with

the hurdles faced in the translation of idioms and phrases and their translation results show the inaccuracy and impreciseness.

In order to convey a similar but not exactly the same meaning, people of various languages use different fixed expressions in such a fashion that while a phrase might seem easy and simple to understand for the ones who represent it, the same combination of words may seem completely dim and unclear. This reflects that every language has got some culture-specific components that are completely different from the corresponding items existing in another language. This leads us to the need of carefully finding the idiomatic pairs in both the languages so that the efficient translation can be done. Hence, there exist two important concerns in this scenario:

1) How to understand the semantics and meanings of idioms and phrases of a specific language.

2) How to align and recreate the same sets of idioms and phrases of source language to target language in a style that they might express exactly the same meaning as of the original language. Parallel corpus for idioms and phrases is better solution for the problem just described above.

IX. METHODOLOGY

We are to develop a parallel corpus for idioms and phrases in English and Punjabi. English and Punjabi language has a vast collection of idioms and phrases. Both English and Punjabi languages are quite different and therefore it is not necessary that both will be having equivalency in multi word expression semantics. We need the database of idioms and phrases for both the languages involved in the matching process. We need the bilingual English-Punjabi dictionary as a resource for accomplishing our work.

Steps involved in the Research are as follows:

1. Collection of idioms and phrases.
2. Collection of bilingual dictionary and Synonyms.
3. Alignment and matching technique used with the meanings of the idioms and phrases.
4. Development of Parallel Corpus for idioms and phrases.
5. Analyze and Test the efficiency of Alignment technique by applying it on different scripts.

Word level and Phrase level alignment technique is used for the matching of equivalent expressions of both the languages. The alignment is done on the basis of meanings and semantics of the idioms and phrases. The matching Experiment uses the type ‘adjective’ of the words from the tokenized set of expressions to identify the context and intentions of the idioms and phrases. Bilingual English-Punjabi dictionary has been used to fetch the meanings of the tokenized words.

X. RESULTS AND CONCLUSIONS

We performed the alignment/matching experiment and recorded the set of idioms and phrases which aligned to their target language equivalents and prepared the parallel corpus for the same. Analyzing the recorded set of matching with their real equivalents, we found the correct and wrong mappings to ensure the accuracy of alignment technique. After performing the alignment experiment based on the word level matching, it has been found that there are total of 1300 matches those clicked to the idea of using an adjective of the word and its effect on the adjacent token. Out of 1300 matches, some of the mappings were one-to-one which means that for each of the expression, there exists one corresponding match in the target language while there are other expressions for which there exists more than one mapping in the target language which serve as probabilistic translation dictionary in the machine translation process.

Some of the correct mappings of the experiment are shown below:

English: A cushy job.

Punjabi: ਖਬੇ ਹੱਥ ਦਾ ਕੰਮ ਹੋਣਾ |

English: A one-tracked mind.

Punjabi: ਇਕੋ ਰਾਗ ਗਾਈ ਜਾਣਾ |

English: Crying shame.

Punjabi: ਅੱਖਾਂ ਲੁਕਾਉਣਾ |

English: Doesn’t know which end is up.

Punjabi: ਅਕਲ ਗਿੱਟਿਆਂ ਵਿਚ ਹੋਣਾ |

Some of the Wrong mappings of the experiment are shown below:

English: A bad omen.

Punjabi: ਮੱਖੀ ਨਿਗਲਣੀ |

The above wrong match is due to the inability to characterize the intentions as ‘A bad Omen’ means a sign or indication that something bad is going to happen but the alignment fails to catch the intentions

and gives the results only for an act of doing something bad.

English: Back down.

Punjabi: ਪੁੱਠੀ ਪੱਟੀ ਪੜ੍ਹਾਉਣਾ |

The above wrong match is due to the ambiguity of words matched, as 'Back down' means reversing your opinion and admit defeat but the meaning of reverse opinion in Punjabi is ਉਲਟੀ ਮੱਤ and hence resulted in the wrong mapping.

The ambiguity problems and the inaccuracy to identify the intentions of the multi word expressions resulted in the wrong mappings from source to the target language. The results show us that there many words having more than one dictionary meaning, which does not contribute to the intentions of the idioms and phrases but still got matched with the tokens in the target language and hence resulting in a wrong mapping.

XI. REFERENCES

[1]. Krings, H.P, "Translation problems and translation strategies of advanced German learners of French" , in the precedings of Interlingual and intercultural communication ,pp. 263-75,1986.

[2]. Gurpreet Singh Josan and Monika Gaule, "Machine Translation of Idioms from English to Hindi", in International Journal Of Computational Engineering Research, vol.2, pp. 5-54, Oct-2012.

[3]. Amir Shojaei " Translation of Idioms and Fixed Expressions: Strategies and Difficulties", in the proceedings of Theory and Practice in Language Studies,vol.2, pp. 1220-1229, June-2012.

[4]. Sofia Trypanagnostopoulou, Janet DeCesaris, "Using a Parallel corpus as a dictionary Resource: Studying Idioms in an English-Greek Parallel Corpus", in the Proceedings of 9th conference on Hellenic language and Terminology, pp.211-220, Nov-2013.

[5]. Xiaoping Jiang and Josta van Rij-Heyligers, "Parallel Corpus in Translation Studies: An Intercultural Approach" in the international symposium on Using Corpora in Contrastive and Translation Studies, pp.1-27, September-2008.

[6]. Linli Chen,"Integrated Translation Approach of English Idioms", in the Journal of Language Teaching and Research, Vol. 1, No. 3, pp. 227-230,2010.

[7]. S.K. Dwivedi and P. P. Sukadeve, "Machine Translation System Indian Perspectives", Proceeding of Journal of Computer Science Vol. 6 No. 10. pp 1082-1087,2010.

[8]. AminehAdelnia, HosseinVahidDastjerdi, "Translation of Idioms: A Hard Task for the Translator", English Department, University of Isfahan, Isfahan, Iran ,Theory and Practice in Language Studies, Vol. 1, No. 7, pp. 879-883,2011.

[9]. M. Baker, "A coursebook on translation" London and New York: Routledge,1992.

[10]. Margarita Straksien," Analysis of Idiom Translation Strategies from English into Lithuanian", studies about languages, Vol No.14, pp.13-19,2009.