IJCSC International Journal of Computer Science and Communication Vol. 2, No. 2, July-December 2011, pp. 611-613

EXTRACTING PROVERBS IN MACHINE TRANSLATION FROM HINDI TO PUNJABI USING RELATIONAL DATA APPROACH

Monika Sharma[1] and Vishal Goyal[2]

M.Tech. (ICT) Student, Assistant Professor,
Department of Computer Science, Punjabi University, Patiala
E-mail: sharma.monika95@gmail.com [1], Vishal.pup @gmail.com [2]

ABSTRACT

Proverb is a group of two or more words which cannot be directly translated into another language word by word. These groups of words have a special behaviour. Machine Translation faces a lot of complex problems from its origination. Extracting proverbs is also one of the complex problems in Machine Translation. Finding proverbs during translating a sentence from Hindi to Punjabi, through existing solutions, takes a lot of time. We have designed a simple relational data approach to find and handle proverbs. This approach handles proverbs efficiently.

1. INTRODUCTION

Machine Translation (MT) is the application of the computers to the task of translating texts from one natural language to another. Machine Translation also known as "Automatic Translation" or "Mechanical Translation" is the name for computerized methods that automate all or part of the process of translating from one human language to another. Machine translation is a computer application to the task of analyzing the source text in one human language and producing an equivalent text called 'translated text' or 'target text' in the other language with or without human assistance as it may require a pre-editing and a post-editing phase.

Proverb is a term used to refer to such words which are combination of two or more words. The meaning of a proverb is not a straightforward composition of the meanings of its parts. Meaning is completely different from the free combination.

1.1 Hindi Language

Hindi is one of the major languages of India. It is the 5th most spoken language in the world with more than 180 million native speakers. It is written in the Devanagari script. It is the national language of India and is the world's second most spoken language.

1.2 Punjabi Language

Punjabi is the official language. It is mainly spoken in Punjab and neighbouring states of Haryana, Himachal Pradesh and Delhi. More than 90 million people speak Punjabi language and it is one of the top 10 languages in the world by number of speakers. It is written in Gurmukhi and Shahmukhi scripts.

2. RELATED WORK

Multiword expression refers to combination of more than one word with same or different part of speech structure and giving single meaning. The main sources of multiword expressions are phrases and idioms. Other than idioms and phrases, other words also act as multiword expressions when combined together.

For example, the word "foreign minister" is a multiword expression, which is formed by the combination of two different words "foreign" and "minister".

Segond and Tapanainen's work on 'Using a Finite-state based Formalism to Identify and Generate Multi-word Expressions' (Segond & Tapanainen,1996) demonstrates how a multiword expression can be encoded, and how their compiler would use them to identify the MWEs.

Wehrli's work on translating idioms (Wehrli, 1998) talks about how MWEs can be used by a linguistic system. It also talks about the transfer and generation of idioms in its framework.

Vineet Kumar Birla, Mohd. Nabeel Ahmed, V. N. Shukla (2009) presented an algorithm for extraction of Multiword Expression from a given English text. To find out the appropriate words for the lexical database, they used the extraction process for multiwords expression and unique words etc. The extraction process is comprised of multiple algorithms used at different phases of the process. They described Extraction Methodology used in Multiword Expression in which there are five steps: Text Corpora Cleaning, Tagging, Parsing, File processing and Extractor. Then they implemented methodology for Anglabharati machine translation system. It has proved its simplicity in the extraction

process of Multiword expressions and unique words with practical application.

3. PROBLEM DEFINITION

Proverbs are known to constitute a serious problem for natural language processing (NLP). Resolving proverbs into their correct meaning in MT is a tedious job. The problem domain, to which this paper is concerned, is finding and translating Hindi Proverbs into Punjabi during translation process. During Hindi into Punjabi translation process, we translate each Hindi proverb into Punjabi proverb word by word. But there are some combinations of two or more words which cannot be directly translated into Punjabi word by word, such combination of two or more words are called proverbs and need to be translated into some particular word or combination of words. This is very rare that Proverbs are present in the input text for MT System but there is a need to extract proverbs from input text and translate them into correctly. My problem is to design a Graphical User Interface, which accepts input as a Hindi Language word (source text) from the keyboard and converts it into Punjabi Language word (target text). The source text is converted into target text in Unicode Format.

For example:

Sentence in Hindi:

जून के महीने में अंगार बरसते हैं।

Punjabi text becomes:

ਜੂਨ ਦੇ ਮਹੀਨੇ ਵਿੱਚ **ਅੰਗਾਰ ਵਰ੍ਹਦੇ** ਹਨ।

But this is not correct translation.

The correct translation is:

ਜੂਨ ਦੇ ਮਹੀਨੇ ਵਿੱਚ ਬਹੁਤ ਗਰਮੀ ਪੈਂਦੀ ਹੈ।

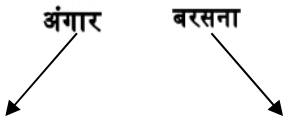
4. METHODOLOGY

The identification and extraction of proverbs is a problem more complex than can be dealt with by one simple solution. The choice of approach depends on the nature of the task and the type of the resources used. We discuss the experiment we conducted to extract and validate proverbs.

4.1 Relational Data Approach

To resolve proverbs, we have used relational data approach. We have created a database in which Hindi proverbs with Punjabi meanings are stored. We have divided Hindi and Punjabi proverbs into two parts: Static and Dynamic.

Example: अंगार बरसना



Static Part Dynamic Part

Static part is translated by Regular Expression.

Inflections of this dynamic part can be:

बरसते

बरस रहे

बरसने

Regular expression provides a concise and flexible means for matching strings of text, such as particular characters, words, or patterns of characters. It describes a set of strings. It is usually used to give a concise description of a set, without having to list all elements.

4.2 Method

First, we extract Hindi proverb from the sentence and then static part of the proverb is handled by using regular expression. It checks the proverb with the database. In case it finds the proverb then it gets the equivalent Punjabi meaning of the proverb. This provides us with pre-defined knowledge of what concepts are likely to be represented as proverb in Punjabi language. Then, we translate the Hindi proverb into Punjabi. Finally, we validate the result.

5. RESULTS

We have tested this method on more than 800 proverbs. Using the Relational Data Approach we have reported an accuracy of 60-80%.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented the technique for finding and translating Hindi proverbs into Punjabi during translation process. Using the Relational Data Approach we have reported an accuracy of 60-80%. As future work, database can be extended to include more proverbs to improve the accuracy. This technique can be used as a teaching aid to all the students from Class-V to the highest level of education. By adding new more features, we can upgrade it to learn all the aspects of Hindi Grammar. It can also be used to solve and test the problems related to Hindi Grammar.

ACKNOWLEDGEMENT

We would like to thank Dr. G.S. Lehal, Professor, Department of Computer Science, Punjabi University, Patiala for many helpful suggestions and comments.

REFERENCES

[1] Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith (2010), "Automatic Extraction of Arabic Multiword Expressions", *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pp. 18-26, Beijing, August 2010.

[2] Vineet Kumar Birla, Mohd. Nabeel Ahmed, V. N. Shukla (2009), "Multiword Expression Extraction – Text Processing", *Proceedings of ASCNT – 2009*, CDAC, Noida, India, pp. 72-77.

[3] Vishal Goyal and Priyanka (2005), " Implementation of Rule Based Algorithm for Sandhi-Vicheda of Compound Hindi Words", *International Journal of Computer Science Issues*, **3**(2009), pp. 45-49.

[4] Kashif Bilal, Uzair Muhammad, Atif Khan, and M. Nasir Khan (2005), "Extracting Multiword Expressions in Machine Translation from English to Urdu using Relational Data Approach", *World Academy of Science, Engineering and Technology*, **12**, 2005, pp. 1-3.

[5] Eric Wehrli (1998), "Translating Idioms", Proceeding COLING '98 Proceedings of the 17th International Conference on Computational Linguistics - Volume 2 ©1998. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2 ©1998, pp 1388-1392.