IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Identification and Translation of Idiomatic Sentence from Hindi to English

N. Thakre[1*], V. Gupta[2], N. Joshi[3]

[1]IES, IPS Academy, RGPV, Indore, Madhya Pradesh, India
[2,3] Apaji Institute, Banasthali University, Rajasthan, India

[*]*Corresponding Author:* [1]*nainathakre2@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 24/May/2018, Published: 31/May/2018

***Abstract* —** This investigation is designed to shed light on identification and translation of idiomatic expression from Hindi to English language. The main problem encountered in idiomatic translation was investigated. Identification of idiom is utmost important resource for machine translation system and research. Machine translation system has been developed for many languages such as we have google translator, Bing translator etc. But they are failed to provide the correct translation of sentences containing idioms. The idiom parallel corpus was manually created to test the generated resource. The sentences containing idioms are translated with google translate system and noticed that it does not provide figurative meaning [1] which makes the translation of idioms rather difficult than any text translation. They are the real challenge for machine translation from the preliminary stage of machine translation development. A lot of research has been done for extraction and translation of text in many languages, but no significant research has been captured in Hindi to English idiom translation. In this paper we have given a rule based approach for the identification and translation of idiom using machine translation. The aim of a proper idiom translation is achieving equivalent sense and provide figurative meaning, strategies, cultural aspects and effects. The output is evaluated manually for intelligibility and accuracy. Further This Hindi to English idiom translation system can be expanded for other language pairs to improve their translation by encapsulating correct idiom translation with their ordinary translation.

***Keywords*—**Idioms, Idiom translation, Idiom identification, Machine translation, Hindi, English, Language.

I. INTRODUCTION

Machine Translation automatically converts one natural language to another through computer. It converts the text of source language (SL) preserving its meaning to the fluent text in target language (TL). Machine translation is a computer application for converting the text from one language to other with or without human assistance as it may require a pre-editing and a post-editing phase. An Idiom is phase or expression that has a figurative meaning. An idiom's figurative meaning is different from the literal meaning [2]. Meaning of an idiom is not predictable from usual meaning of its constituent elements. Every language has its own set of idioms, English and Hindi are abundant in idioms. Idioms are important part of conversation and are frequently used in wide variety of situations from friendly conversations to business and more formal and written context. An idiomatic expression may convey a different meaning, that what is evident from its words that's why they are hard to translate into other language. Therefore, it is important to identify the idioms and then replace them with the suitable and idiom carry appropriate meaning in any other language while translation. This is still an important topic for research and serious drawback of many machine translators like google. For example, consider following snapshots taken from google translator with input given as Hindi idiomatic sentence and converted to English.

Fig. 1: Example 1 Google snapshot.

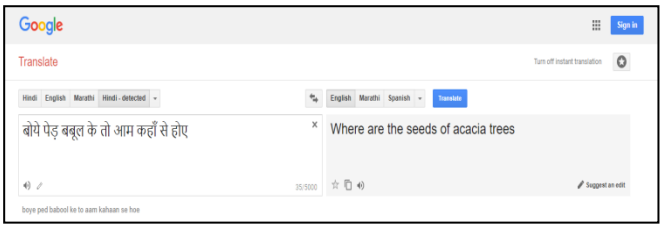

Fig. 2: Example 2 Google snapshot

Clearly, the output does not convey the intended meaning in target language.

II. LINGUISTIC BACKGROUND

A. Hindi

Hindi language is derived from Hindustani which is derived from Sanskrit language. It contains much vocabulary from Sanskrit language and is also written as such. It is the official language of India, where majority of population communicates using this language and it is 4th most spoken language in the world. It used to be written in Brahmi script but now it is written in Devanagari script. Devanagari consists of 11 vowels and 33 consonants and is written from left to right. Unlike Sanskrit, Devanagari is not entirely phonetic for Hindi, especially failing to mark schwa dropping in spoken Standard Hindi.

B. English

English is a West Germanic language, and 3rd most spoken language of the World. It is closely related to Frisian languages but vocabulary is influenced by other Germanic languages. There are noticeable variations among the accents and dialects used in different countries. It is mostly analytic pattern with little inflection, a fairly fixed SVO word order and a complex syntax.

III. LITERATURE SURVEY

In this section the main focus is on the work done in the Indian context instead of discussing idiomatic translation. Idiomatic Machine Translation efforts in Indian context dated to October 2012 with Gaule & Josanal [3] in their research the investigation is designed to shed light on the identification and translation of idiomatic expressions from English to Hindi is analysed. They gave different strategies of idiom in machine translation Using an idiom of similar meaning and form, Using an idiom of similar meaning but dissimilar form, Using an Idiom Translation by paraphrase, Using an Idiom Translation by Omission and Online MT Systems There are following MT systems that have been developed for various natural language pair. Systran is a rule based Machine Translation System developed by the company named Systran. It was founded by Toma in 1968. It offers translation in about 35 languages. It provides technology for Yahoo! Babel Fish and it was used by Google till 2007. Design overview- identification of idiom and process them and the processed sentence will be used as input by translation system. The result for evaluation were-30% sentences were correctly translated by sentences directly obtained from goggle translation system and 70% sentences were correctly translated by sentences pre-processed by our system and then translated from goggle translation. Further with Rajesh Kumar Chakrawarti, Himani Mishra, Dr. Pratosh Bansal [4] they witnessed several significant advancements in Natural Language Processing , which has let text and speech processing to make huge gateway to world-wide information source [5]. The paper focuses on the techniques and approaches like corpus-based, rule-based, direct and hybrid approach. used for machine translation systems together with their example systems. They identified the problem of Structural Divergences, Approach used and ambiguity, cultural problem and named entity. There were several limitations identified such as. 1)Dictionary used: -Translation will be greatly affected by the depth and richness of dictionary used. 2) As it is a machine, failure of the machine can't be predicted. Like all other system, it may crash down at any instance. 3) Idioms are difficult to interpret as they point to some other meaning than the words used. In this survey paper, they studied various MT approaches, techniques, and many machine translation systems together with their benefits and limitations in a longitudinal and latitudinal way.

Further, approaches of idiom translation are described by Rajesh Kumar Chakrawarti, Himani Mishra, Dr. Pratosh Bansal [6]. This research paper proposes a different system architecture for idiom translation from Hindi to English. This architecture resembles a Rule-based approach in which Transfer-based method is used which converts the idioms having similar meaning and either similar or dissimilar form from Hindi to Tokens and then in English and Interlingua based method is also used which converts the typical idioms whose meaning is not given as it is in the used database to their simple meanings and then they are converted in English. This architecture contains two phases

Phase I- Comparison phase, in which the input is compared to the database

Phase II- Translational phase, in which the translation happens.

This work is a modified version of Machine translational system which can be embedded with other machine translational systems to get better results. Further Survey of machine translation system in India is given by Garje and Kharate [7]. They focused on different approaches used in the development of Machine Translation Systems and also briefly described some of the Machine Translation Systems along with their features, domains and limitations. They gave brief history of machine translator system at International level staring from 1948 to 2010. They listed machine translation systems and various approaches used in developing these systems.

1. Direct Machine Translation Systems - Anusaaraka systems among Indian Languages (1995), Punjabi to Hindi MT System (2007, 2008), Web based Hindi-to-Punjabi MT System (2010), Hindi-to-Punjabi MT System (2009, 2011).
2. Transfer-Based MT Systems - Mantra MT (1997), MANTRA MT(1999), An English–Hindi Translation System (2002), MAT (2002), Shakti (2003), English-Telugu MT System (2004), Telugu-Tamil MT System (2004), OMTrans

(2004), The MaTra System (2004, 2006), English-Kannada machine-aided translation system (2009), Tamil-Hindi Machine-Aided Translation system (2009), Sampark System: Automated Translation among Indian Languages (2009).
3. Interlingua Machine Translation Systems - ANGLABHARTI (2001), UNL-based English-Hindi MT System (2001), AnglaHindi (2003).
4. Hybrid Machine Translation Systems
5. Example Based Machine Translation (EBMT) Systems
6. Statistical Machine Translation Systems.

In this paper author described MT techniques in a longitudinal and latitudinal way with an emphasis on the MT development for Indian languages as well as non-Indian languages and concluded that almost all existing Indian language MT systems are based on rule-based, hybrid and statistical approaches. D. Brar & R. Kaur [8] discussed the problems associated with the idiomatic translation. They presented the definition of idioms to see what they are. Then, classified the idioms into different categories and in the end, gives some techniques and procedures to translate them. They grouped idioms into five categories of colloquialisms, proverbs, slang, allusions and phrasal verbs. And concluded that idioms are arguably the most complex and problematic task for translators.

IV. METHODOLOGY

For the better translation of idioms from Hindi to English, a custom algorithm is used. This algorithm requires the input file of list of all idioms and their corresponding English idioms rather than their word to word conversion unlike google translate.
It then checks the sentence given as input by the user for a matching idiom in the file and replaces with its English counterpart, then translates the sentence, resulting in appropriate replacement of the idiom.
This algorithm works in steps, as follow:

A. Collection of parallel idiom corpus

Parallel Database of 1000 Hindi and English idiom is created manually from different source and a text file is created.
A text file is provided to the program as a collection of idioms in Hindi and English. Each line of this file stores Hindi idiom and its English counterpart, separated by a hyphen. A single line can only have one hyphen present. The program will read the file, line by line making two lists on idioms. First list will store Hindi idioms and the second list contains matching English idioms at the same indices.

B. Idiom identification

Input is taken as a Hindi sentence containing an idiom. This sentence is then split into word stored in a list. This list is matched to all the idioms in the list, word by word. At the end of the loop, a final idiom is identified.

C. Idiom replacement

Once the idiom is identified, Using the index of the final idiom in first list, the idiom in the second list is extracted, and the idiom in the sentence is replaced by it.

D. Translation of idiomatic sentence

Once the idiom is replaced in the sentence, the sentence is then translated to English using *mtranslate* module for python, which works as Google Translator. Since, the final sentence contains Hindi words and the correct English translated idiom, the output sentence by the module considers the idiom and put it as it is, giving a properly translated idiomatic sentence.

E. GUI

A GUI (Graphical User Interface) developed in the same language is used for ease of input and output, so that the user doesn't have to run the program from command line again and again.

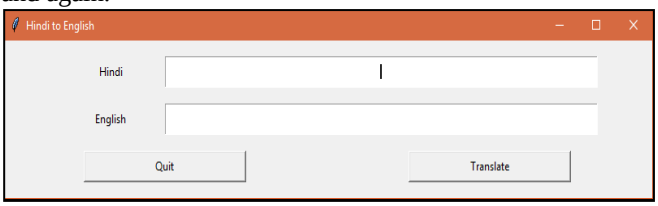

Fig.3. *GUI Snapshot*

F. Flowchart

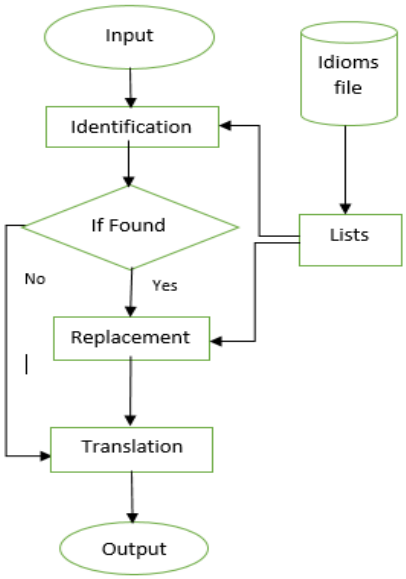

Fig 4: Flowchart of overall method

V. RESULTS AND DISCUSSION

Output is counter checked manually whether the sentences are translated perfectly or not. Some snapshots we obtained from the proposed technique and google translate for comparison. The result for the evaluation is that 86% sentences were correctly translated by the system designed and Google provide 30% correct translations for the idiomatic sentence. The accuracy is checked for 50 sentence calculated by the equation 1 given below-

Accuracy = (Correct output ÷ Total Input) × 100 % (1)
= (43 ÷ 50) × 100
= 86 %

Fig 5: *Google Translator snapshot 1.*

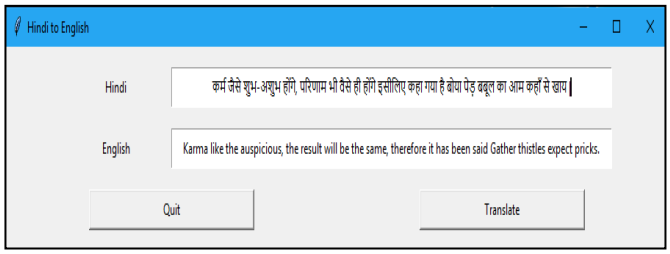

Fig 6: *Proposed GUI snapshot 1*

Fig 7: *Google Translator snapshot 2.*

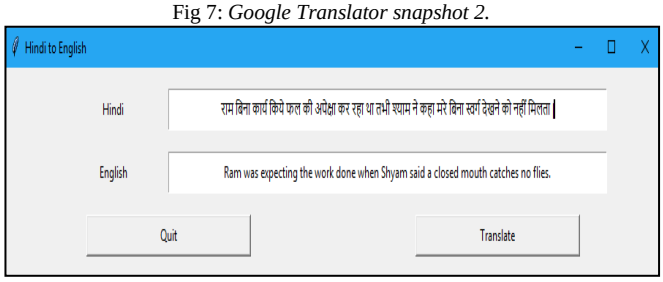

Fig 8: *Proposed GUI snapshot2.*

VI. CONCLUSION AND FUTURE SCOPE

The paper presented the technique for identifying and translating Hindi idiomatic sentences to English. The rule based and machine translation approach for Identification and Translation is proposed. Testing data is manually created to test the generated resource. The output is evaluated manually for intelligibility accuracy. 86% accuracy is reported Analysis of results shows that the problems of bad translation are due to errors of and different categories like-irrelevant idioms and grammar etc. So from this evaluation experiment the identification and translation of idiomatic sentence form machine translation will increase its accuracy from existing system. There are various application in machine translation, information retrieval & language processing etc. As future work, database of idioms can be extended to include more idioms to improve the accuracy.

REFERENCES

[1] D. Anastasiou, "*Idiom Treatment Experiments in Machine Translation"*, Cambridge Scholars Publishing, 2010.

[2] The Oxford companion to the English language (1992:495f.).

[3] M. Gaule & Josan, G. S., "*Machine Translation of Idioms from English to Hindi"*, International Journal Of Computational Engineering Research, *2*(6), 2012.

[4] R. K. Chakrawarti, H. Mishra, & P. Bansal, "*Review of Machine Translation Techniques for Idea of Hindi to English Idiom Translation"*, International Journal of Computational Intelligence Research, 13(5), 1059-1071, 2017.

[5] F. Ciravegna, S. Harabagiu, "*Recent Advances in Natural Language Processing*", IEEE magazine, computer.org/intelligent, 2013.

[6] H. Mishra, R. K. Chakrawarti & P. Bansal, "A New Approach for Hindi to English Translation", International Journal on Computer Science and Engineering, Vol. 9 No.07, 0975-3397, Jul 2017.

[7] G. V. Garje & G. K. Kharate, "*Survey of machine translation systems in India*", International Journal on Natural Language Computing (IJNLC) Vol, 2, 47-67, 2013.

[8] D. Brar & R. Kaur, "A Review of Transliteration system from English to Punjabi", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4 (7), 2277 128X, July 2014.

[9] V. Gupta, N. Joshi & I. Mathur, "Approach for Multiword Expressions and Recognition & Annotation in Urdu Corpora", Image Information Processing (ICIIP), Fourth International Conference, 2017. IEEE 2017.

[10] V. Gupta, N. Joshi & I. Mathur, "Design & Development of Rule Based Inflectional and Derivational Urdu Stemmer 'Usal'", INBUSH-ERA-2015, 7-12, 2015. IEEE 2015.

[11] V. Gupta, N. Joshi & I. Mathur, "Rule Based Stemmer in Urdu", Computer and Communication Technology (ICCCT), Fourth International Conference, 2013. IEEE 2013.

Authors Profile

Naina Thakre is pursuing her Bacherlor of Engineering from Institute of Engineering and Science, IPS Academy, Indore from Computer Science Department, currently in 3rd year.

Vaishali Gupta is pursuing her Ph.D in Computer Science & Engineering from Banasthali University, Rajasthan, India. She has

interest in language processing specifically for Indian Languages. She has developed various NLP tools for Hindi and Urdu language. Her current research interest includes Natural language processing, Machine Translation and Information retrieval.

Nisheeth Joshi works as an Associate Professor at Banasthali University. His areas of interest include computational linguistics, Natural Language Processing, and artificial intelligence. Besides this, he is also very actively involved in the development of MT engines for English to Indian languages. He is one of the experts empaneled with the TDIL program, Department of Information Technology, Govt. of India, a premier organization that oversees Language Technology Funding and Research in India. He has several publications in various journals and conferences and also serves on the program committees and editorial boards of several conferences and journals.

