# Machine Translation of Idioms from English to Hindi

**Monika Gaule[1] Dr. Gurpreet Singh Josan[2]**

[1]M.Tech. (ICT), Department of Computer Science, Punjabi University, Patiala, India
[2] Assistant Professor ,Department of Computer Science, Punjabi University, Patiala, India

## Abstract

This qualitative investigation is designed to shed light on the identification and translation of idiomatic expressions from English to Hindi is analyzed in this paper. It investigates the main problems and difficulties encountered during idioms translation and the extent to which idiomaticity is retained.This identification of idioms is utmost important resource for machine translation system and research in this field. A rule based approach for identification of idioms is proposed. The sentences containing idioms are translated with goggle translate system. We have manually created testing data to test the generated resource.The aim of a proper idiom translation is achieving equivalent sense, strategies, cultural aspects and effects. The output is evaluated manually for intelligibility and accuracy. The accuracy of system is 70%. Analysis of results shows that the problems of bad translation are due to errors of different categories like-irrelevant idioms, grammar agreement, part of speech etc.

**Keywords** –idiom,idiom identification, idiom translation,translation strategies

## 1. “Introduction”

Machine translation (MT) is defined as the use of a computer to translate a text from one natural language, the source language (SL), into another one, the target language (TL), Machine translation is a computer application to the task of analyzing thesource text in one human language and producing an equivalent text called ‘translated text’ or ‘target text’ inthe other language with or without human assistance as it may require a pre-editing and a post-editing phase. Every language has its own idioms, a special kind of set expressions that have developed within a language. English and Hindi are abundant in idioms. One of the most important aspects of English is idioms. They are frequently used in a wide variety of situations, from friendly conversations and business meetings to more formal and written contexts. An idiom is a group of words which has, as a whole, a different meaning from the meaning of its constituents.In other words, the meaning of the idiomatic expression is not the sum of the words taken individually. Idioms are fairly controversial. There is no one set definition of what an idiom is. The word itself comes either from Latin idioma, where it denotes special property, or from Greek idiōma, meaning special feature, special phrasing. Hence, the logic imposes associations with elements of language phrases that are typical for a given language and, therefore, hard to translate into another language. An idiomatic expression may convey a different meaning, that what is evident from its words. For example English: It’s raining like cats and dogs

Hindi translation By Google*:* अपनी बिल्लियों और कुत्तों की तरह बारिश हो रही

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

### English Language

English is now the most widely used language in the world it is the third most common native language in the world, with more than 380 million native speakers. English Language is written in Roman script. It is a West Germanic language that arose in the Anglo-Saxon kingdoms of England. It is one of six official languages of the United Nations. India is one of the countries where English is spoken as a second language.

### 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 second most spoken language

## 2. “Translation Problems”

A translation problem is any type of difficulty in the source language (SL) text that obliges the translator to stop translating. This difficulty is mainly due to grammatical, cultural or lexical problems.

### Grammatical Problems

Grammatical problems are the result of complicated SL grammar, different TL grammar or different TL word order. For example, the word order of English and Hindi is not same. English follows SVO scheme while Hindi Follows SOV scheme. Consider following idiom in English: “Add fuel to fire”

Corresponding Hindi sentence is आग में घी का काम करना.

Here in English phrase, the word “fire” is at last position whereas in Hindi its counterpart आग is at first position of the phrase.

**Cultural Problems**

A number of problems may be raised in cross-cultural translation. The greater the gap between the source and target culture, the more serious difficulty would be. Translation between English and Hindi which belong to two different cultures (the Western and the Indian cultures), and which have a different background is a best example of such problems. Cultural problems may include geographical, religious, social and linguistic ones. Hence, the expression “summer’s day” in „Shall I compare thee to a summer’s day‟ will be best translated into Hindi asग्रीष्मऋतुto convey the same meaning.

**Word sense Ambiguity**

This problem occurs when there are multiple interpretation of words or sentence. Among these problems we have:

**Phrase level ambiguity**

Phrase level ambiguity occurs when a phrase can be interpreted in more than one ways. For example theexpression 'spill the beans' may refer to the beans that are actually spilled or idiomatically the phrase may refer to leak out secret information.

**Word level ambiguity**

The word ambiguity conveys the multiple interpretations of words. For example**to bear the lion in his den** As bear have the multiple meanings भालू,कष्टउठाना,फलदेना,ऋक्ष , उत्पन्नकरना, रीछ, लेजाना

**Different Strategies of Idioms in Machine Translation**

The term "strategy" refers to a method employed to translate a given element unit making use of one or more procedures selected on the basis of relevant parameters. presents a series of strategies used by professional translators.

**Using an idiom of similar meaning and form**

It involves using an idiom in the target language which conveys roughly the same meaning as that of the source-language idiom and, in addition consists of equivalent lexical items. Example: to rub salt in wounds जले पर नमक छिड़कना

**Using an idiom of similar meaning but dissimilar form**

Idioms of similar meaning but dissimilar form refer to those having completely different meanings and the occasions in which the idioms are used are not alike as well. ExampleTo sleep like a log घोड़े बेच कर सोना

**Using an Idiom Translation by paraphrase**

where the expression is often rewritten using other words to simplify it and then translate. Example

The suspension system has been fully uprated to take rough terrain in its stride. निलंबन प्रणाली पूरी तरह से अपनी प्रगति में किसी न किसी इलाके लेने के लिए अद्यत न किया गया है.

And The capacity of the suspension system has been raised so as to overcome the roughness of the terrain. निलंबन प्रणाली की क्षमता को इतनी के रूप में इलाके का खुरदरापन पर काबू पाने के लिए उठाया गया है.

Second example is more intelligible in target language.

**Using an Idiom Translation by Omission**

If the idiom has no close match, the system can simply omit the idiom in target language. The meaning will not be harmed, if this technique is used when the words which will be omitted are not vital to the development of the text. Translators can simply omit translating the word or the expression in question

**Online MT Systems**

There are following MT systems that have been developed for various natural language pair.

**Systran**Systran is a rule based Machine Translation System developed by the company named Systran. It was founded by Dr. Peter Toma in 1968. It offers translation in about 35 languages. It provides technology for Yahoo! Babel Fish and it was used by Googletill 2007 .

**Bing Translator**

Bing Translator is a service provided by Microsoft, which was previously known as Live Search Translator and Windows Live Translator. It is based on Statistical Machine Translation approach.Bing Translator offers 32 languages in both directions.When a URL is introduced, Bing Translator opens a new window showing the text in English and the progress of the translation. Once it is finished it shows the translated webpage. Four different views can be selected

"Side by Side", "Top, Bottom", "and Original with hover translation "and" Translationwith hover original". When the user runs the mouse over one target sentence, it highlights this sentence and the corresponding source sentence.

**Google Translate**

Google Translate is a free translation service that provides instant translations between 57 different languages. Google Translate generates a translation by looking for patterns in hundreds of millions of documents to help decide on the best translation. By detecting patterns in documents that have already been translated by human translators, Google Translate makes guesses as to what an appropriate translation should be. This process of seeking patterns in large amounts of text is called "statistical machine translation". It can translate text, documents, web pages etc. English to Hindi Machine Translation system(http://translate.google.com/), In 2007, Franz-Josef Och applied the statistical machine translation approach for Google Translate from English to other Languages and vice versa, Thus statistical machine translation approach for identification of idioms is proposed.Many online machine translation system are available on internet as no single translation engine will be consistently most effective for all pairs of languages and text conditions . As further we use Google Translate system for translating English Text to Hindi.The accuracy is good enough to word understand the translated text, but not perfect. The system has been available online for use.

**Motivation**

English to Hindi translation system available online at http://translate.google.com/ which translates English text into Hindi text does not extract idioms from the input text during translation. Though, this is very rare that Idioms are present in the input text for MT System but there is a need to extract Idioms from input text and translate them correctly. So we developed an algorithm for finding and translating English Idioms present in the input text and translate them into Hindi text.

**Example:**

Sentence in English: He has **settled** his **account** with me
Output for this by direct goggle translateshttp://translate.google.com/is:
वह मेरे साथ अपने खाते में बसे है
Clearly, the output is not intelligible. But if we somehow, find and replace the idioms in above sentence as follow
S: He has चुकता किया हुआ his हिसाब किताब with me and translate it with goggle translate system we get:
वह चुकता किया हुआ उसके हिसाब किताब मेरे साथ है
which is quite intelligible and much better than previous output and thus motivate us to work in this direction.

## 3. "Design Overview"

Here, the aim is to design a system for identifying idioms and process them. This processed sentence will be then used as input by translation system. The system architecture is as follow

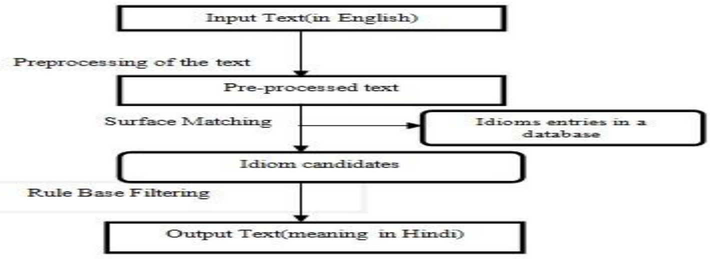
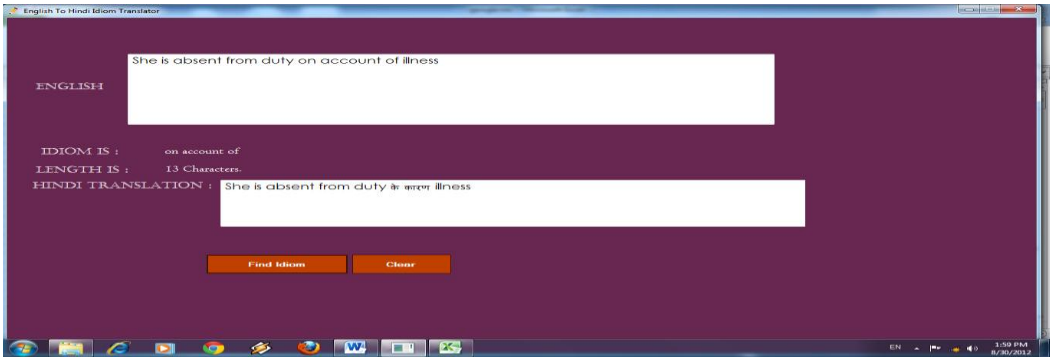


Figure 3.1 the overall flow of the system

The system consists of three modules which includes Preprocessing (Paradigmatic replacement, syntagmatic augmentation, deletion, Replacing inflected form of verbs, Replacing Plural forms of Nouns, articles, personal pronouns), Surface matching (FilteringPart-of -speech tagging and chunking patterns, identifying idiom candidates) and Post processing module.

## Implementation

The system is implemented in Asp.net at front end and MS SQL 2008server at back end. A English to Hindi Idiom Translator System Interface is created whose object will accept a string in English language and returns its corresponding Hindi string. When the user input an English text ,and clicks the "search" button ,the system outputs the idiom candidates with their meaning (in Hindi).Here, we present the use of regular expressions in a translation system for extracting and translating the English idioms. (a)Idiom Extraction using Pattern matching**:** If there is a idiom in sentence, it is extracted from the sentence by using Pattern matching. Search pattern matches the Idiom in the sentence and extracts the Idioms from the sentence.Translating (b)English idioms to Hindi idioms:Now, the extracted Idioms is replaced with the equivalent Hindi meaning of that Idiom. It means English idiom is translated into Hindi idioms.



English To Hindi Idiom Translator

ENGLISH: She is absent from duty on account of illness

IDIOM IS : on account of

LENGTH IS : 13 Characters.

HINDI TRANSLATION : She is absent from duty के कारण illness

**English to Hindi Idiom Translator System Interface**

## 4. “Evaluation and Results”

### Test Data

We carried out evaluation experiments in order to observe the overall performance of the system, as well as the following three aspects: (1) the effect of standardization of words; (2)the effect of the POS-based filtering; (3) the overall performance of the system. To evaluate the machine translation system several methods are used . These evaluation methods can be categorized into groups namely the sentences directly obtained from goggle translation system (Case I) and the sentences preprocessed by our system and then translated from goggle translation (Case II) . We observe how many correct idiom candidates our system was able to locate with each set of data. The data set used for evaluation were (a)100 sentences containing idiom variations, randomly extracted from the idiom variation, (b) data consisting of 20 news ( sports, politics, world) , articles (From the different internet websites), stories (published by various writers) We manually identified and tagged the idioms for these articles.

### Evaluation Metric

The survey was done by 10 People of different professions. All persons were from different professions having knowledge of both English and Hindi Language. Average ratings for the sentences of the individual translations were then summed up (separately according to intelligibility and accuracy) to get the average scores. Percentage of accurate sentences and intelligent sentences is also calculated separately by counting down the number of sentences.Two type of subjective tests will be performed viz. Intelligibility and Accuracy. Intelligibility test which is effected by grammatical errors, mistranslations, and un-translated words.The evaluators are provided with source text along with translated text. A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called accuracy.

### Intelligibility test score

Score3:Idioms that are actually used in input text that is sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.
Score 2: Variations of (1) is created by replacement of articles that is sentence is generally clear and intelligible.
Score 1: Variations of (1) created by singular/ plural forms of nouns that is sentence contains grammatical errors &/ or poor word choice.
Score 0: That is sentence is unintelligible the meaning of sentence is hopeless.

### Accuracy test score

Score 3 : Completely Faithful
Score 2: Fairly faithful: more than 50 % of the original information passes in the translation.
Score 1: Barely faithful: less than 50 % of the original information passes in the translation.
Score 0: Completely Unfaithful. Doesn’t make sense

### Evaluation Method

Thus the output is counter checked manually whether the sentences are translated perfectly or not. Following formula was used to find the accuracy test of case I and case II .

**Accuracy percentage =(Number of correctly sentences/ Total number of sentences )*100**

**Results**

Results from the evaluation are:

- 30% sentences were correctly translated by sentences directly obtained from goggle translation system (Case I)
- 70% sentences were correctly translated by sentences preprocessed by our system and then translated from goggle translation (Case II)

So In this evaluation experiment ,sentences preprocessed by our system andthen translated from goggle translation(case II) is high accuracy and is much better than sentences directly obtained from goggle translation system (Case I)

**Analyzing the results, certain patterns of errors and misses were identified.**

1)Errors resulting from insufficiency of lemmatization by the POS-tagger.For instance, "She horribly damned him with faint praise" is based on the idiom "horribly damn with faint praise". However, our system could not detect his idiom because "damned" was recognized as an adverb rather than the verb "damn". Another example The leader "added fuel to fire" by provoking the angry mob to attack the police van however our system could not detect his idiom because "fuel" was recognized as an fuels (noun plural) rather than the verb fuel .नेता नाराज भीड़ उत्तेजक पुलिस वैन पर हमला करके आग में घी डालने का काम This could be avoided by tagging the input sentence with POS. 2)Misses resulting from input text variations in which long phrases are inserted into the idiom constructions. For instance, the dictionary entry "take apart" was not detected for the input text: "She takes (her daughter-in-law)apart with stinging criticism." Another example The shopkeepers **"**beat black and blue" was not detected for the input text: **"**The shopkeepers beat (the thief) black and blue"दुकानदार चोर काले और नीले रंग हरा In order to deal with this, we need to further our understanding of possible idiom variations.

3) Failures resulting from input text variations in which some translations were not translated correctly for example This comedian will have the audience rolling in the aislesइस हास्य अभिनेता को दर्शकों के aisles में रोलिंग है

## 5. Conclusion& Future Work

In this paper, we presented the technique for finding and translating English idioms into Hindi during translation process. The rule based and statistical machine translation Approach for identification of idioms is proposed.The sentences containing idioms are translated with goggle translate system. We have manually created testing data to test the generated resource. The output is evaluated manually for intelligibility and accuracy. Thus we have reported an accuracy of 70%. Analysis of results shows that the problems of bad translation are due to errors of different categories like-irrelevant idioms, grammar agreement, part of speech etc. Thus by this evaluation experiment ,identification of idioms in machine translation from English to Hindi will increase its accuracy from existing system.As future work,database can be extended to include more idioms to improve the accuracy.

## References

[1] David A. Swinney, Anne Cutler (1979) The Access and Processing of Idiomatic Expressions , Journal of verbal learning and verbal behavior 18, pp. 523-534

[2] Martin Volk,1998*T*he automatic translation of idioms. Machine translation vs. translation memory systems. Permanent URL to this publication http://dx.doi.org/10.5167/uzh-19070 assessment of the state of the art. St. Augustin, 167-192. ISBN 3-928624-71-7

[3] ByongRaeRyu et al., (1999) From to K/E: A Korean- English Machine Translation System based on Idiom Recognition and Fail Softening n Proceedings of Machine Translation Summit VII – "MT in the Great Translation Era" pp.469-475, in collaboration with Youngkil Kim, SanghwaYuh, & Sang- kyu Park (ETRI).

[4 ] M.A.Homeidi ,2004, Arabic translation acrosscultures,Vol No 50, pp. 13-27.

[5] Koichi Takeuchi et al.,(2008) Flexible Automatic Look-up of English Idiom Entries in Dictionaries MT Summit XI,10-14September 2007, Copenhagen, Denmark. Proceedings; pp.451-458

[6] Vishal Goyal and GurpreetsinghLehal, 2009, Evaluation of Hindi to Punjabi Machine Translation System, IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, pp 36-39

[7] Linli Chen Yunnan RTV University, March 2010, On Integrated Translation Approach of English Idioms, Journal of Language Teaching and Research, Vol. 1, No. 3, pp. 227-230

[8] MezmazMeryem 2009-2010 Problems of Idioms in Translation Case Study: First Year Master Permanent URL to this publication bu.umc.edu.dz/theses/anglais/MEZ1146.pdf

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

[10] AminehAdelnia, HosseinVahidDastjerdi, July 2011, "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.

[11] Machine Translation online available http://faculty.ksu.edu.sa/homiedan/Publications/ Machine%20Translation.pdf

[12] "Statistical Machine translation ", [outline.] Available http://en.wikipedia.org/wiki/Statistical_machine_translation

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