# HINDI TO PUNJABI MACHINE TRANSLATION SYSTEM

**Vishal Goyal**
Department of Computer Science
Punjabi University, Patiala,India
vishal.pup@gmail.com

**Gurpreet Singh Lehal**
Department of Computer Science
Punjabi University, Patiala,India
gslehal@gmail.com

## Abstract

Hindi-Punjabi being closely related language pair (Goyal V. and Lehal G.S., 2008) , Hybrid Machine Translation approach has been used for developing Hindi to Punjabi Machine Translation System. Non-availability of lexical resources, spelling variations in the source language text, source text ambiguous words, named entity recognition and collocations are the major challenges faced while developing this syetm. The key activities involved during translation process are preprocessing, translation engine and post processing. Lookup algorithms, pattern matching algorithms etc formed the basis for solving these issues. The system accuracy has been evaluated using intelligibility test, accuracy test and BLEU score. The hybrid syatem is found to perform better than the constituent systems.

***Keywords:*** *Machine Translation, Computational Linguistics, Natural Language Processing, Hindi, Punjabi. Translate Hindi to Punjabi, Closely related languages.*

## 1 Introduction

Machine Translation system is a software designed that essentially takes a text in one language (called the source language), and translates it into another language (called the target language). There are number of approaches for MT like Direct based, Transform based, Interlingua based, Statistical etc. But the choice of approach depends upon the available resources and the kind of languages involved. In general, if the two languages are structurally similar, in particular as regards lexical correspondences, morphology and word order, the case for abstract syntactic analysis seems less convincing. Since the present research work deals with a pair of closely related language i.e. Hindi-Punjabi , thus direct word-to-word translation approach is the obvious choice. As some rule based approach has also been used, thus, Hybrid approach has been adopted for developing the system. An exhaustive survey has already been given for existing machine translations systems developed so far mentioning their accuracies and limitations. (Goyal V. and Lehal G.S., 2009).

## 2 System Architecture

### 2.1 Pre Processing Phase

The preprocessing stage is a collection of operations that are applied on input data to make it processable by the translation engine. In the first phase of Machine Translation system, various activities incorporated include text normalization, replacing collocations and replacing proper nouns.

### 2.2 Text Normalization

The variety in the alphabet, different dialects and influence of foreign languages has resulted in spelling variations of the same word. Such variations sometimes can be treated as errors in writing. (Goyal V. and Lehal G.S., 2010).

### 2.3 Replacing Collocations

After passing the input text through text normalization, the text passes through this Collocation replacement sub phase of Pre-processing phase. Collocation is two or more consecutive words with a special behavior. (Choueka :1988). For example, the collocation उत्तर प्रदेश (*uttar pradēsh)* if translated word to word, will be translated as ਜਵਾਬ ਰਾਜ (*javāb rāj)* but it must be translated as ਉੱਤਰ ਪ੍ਰਦੇਸ਼ (*uttar pradēsh)*. The accuracy of the results for collocation extraction using t-test is not accurate and includes number of such bigrams and trigrams that are not actually collocations. Thus, manually such entries were removed and actual collocations were further extracted. The



*Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 1–6,
Portland, Oregon, USA, 21 June 2011. 

correct corresponding Punjabi translation for each extracted collocation is stored in the collocation table of the database. The collocation table of the database consists of 5000 such entries. In this sub phase, the normalized input text is analyzed. Each collocation in the database found in the input text will be replaced with the Punjabi translation of the corresponding collocation. It is found that when tested on a corpus containing about 1,00,000 words, only 0.001% collocations were found and replaced during the translation.

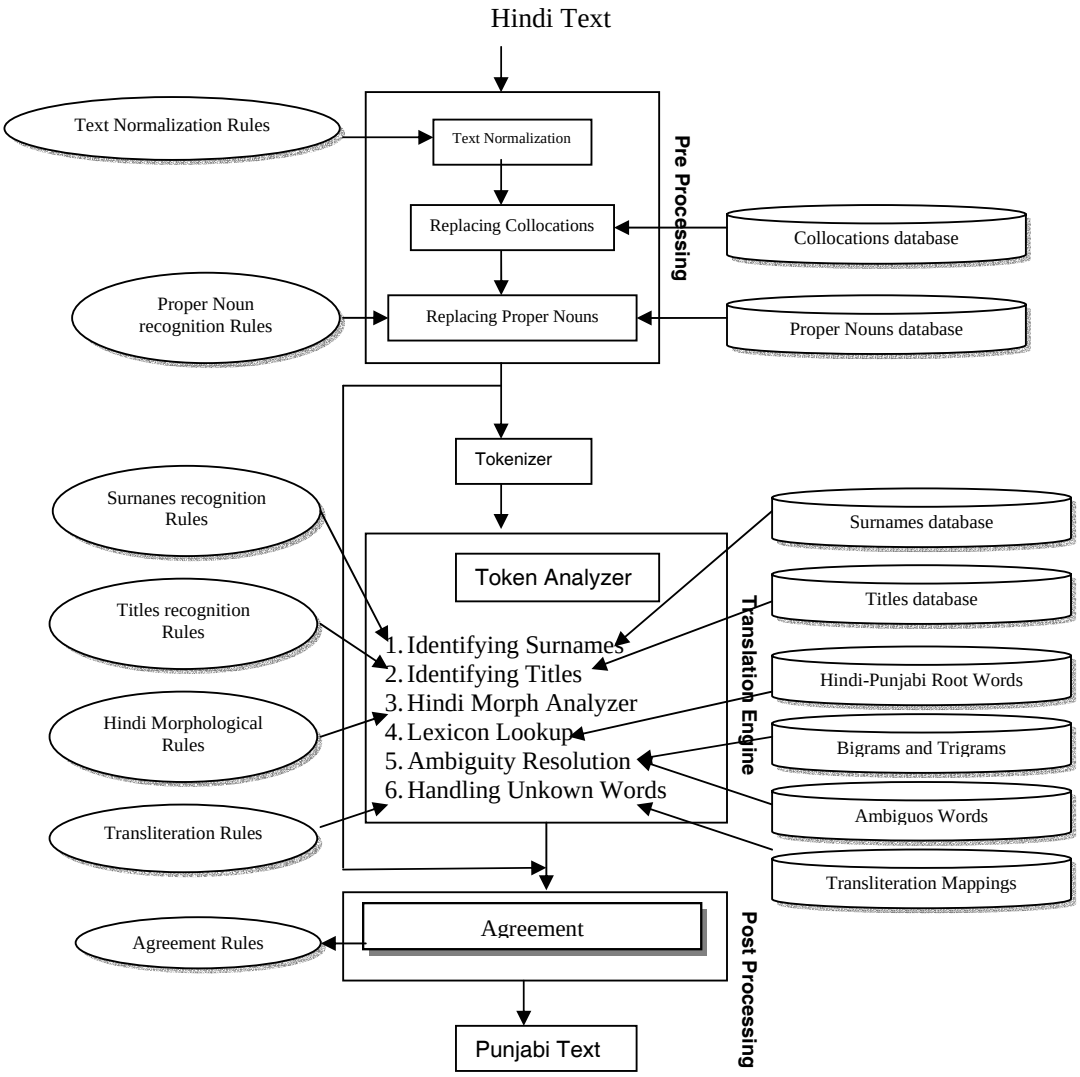


**Figure 1 : Overview of Hindi-Punjabi Machine Translation System**

**2.4 Replacing Proper Nouns**

A great proposition of unseen words includes proper nouns like personal, days of month, days of week, country names, city names, bank names, organization names, ocean names, river names, university names etc. and if translated word to word, their meaning is changed. If the meaning is not affected, even though this step fastens the translation process. Once these words are recognized and stored into the proper noun database, there is no need to decide about their translation or transliteration every time in the case of presence of such words in input text for translation. This gazetteer makes the translation accurate and fast. This list is self growing during each

translation. Thus, to process this sub phase, the system requires a proper noun gazetteer that has been complied offline. For this task, we have developed an offline module to extract proper nouns from the corpus based on some rules. Also, Named Entity recognition module has been developed based on the CRF approach (Sharma R. and Goyal V., 2011b).

### 2.5 Tokenizer

Tokenizers (also known as lexical analyzers or word segmenters) segment a stream of characters into meaningful units called tokens. The tokenizer takes the text generated by pre processing phase as input. Individual words or tokens are extracted and processed to generate its equivalent in the target language. This module, using space, a punctuation mark, as delimiter, extracts tokens (word) one by one from the text and gives it to translation engine for analysis till the complete input text is read and processed.

### 2.6 Translation Engine

The translation engine is the main component of our Machine Translation system. It takes token generated by the tokenizer as input and outputs the translated token in the target language. These translated tokens are concatenated one after another along with the delimiter. Modules included in this phase are explained below one by one.

#### 2.6.1 Identifying Titles and Surnames

Title may be defined as a formal appellation attached to the name of a person or family by virtue of office, rank, hereditary privilege, noble birth, or attainment or used as a mark of respect. Thus word next to title and word previous to surname is usually a proper noun. And sometimes, a word used as proper name of a person has its own meaning in target language. Similarly, Surname may be defined as a name shared in common to identify the members of a family, as distinguished from each member's given name. It is also called family name or last name. When either title or surname is passed through the translation engine, it is translated by the system. This cause the system failure as these proper names should be transliterated instead of translation. For example consider the Hindi sentence श्रीमान हर्ष जी हमारे यहाँ पधारे। (*shrīmān harsh jī hamārē yahāṃ padhārē)*. In this sentence, हर्ष (*harsh)* has the meaning "joy". The equivalent translation of हर्ष (*harsh)* in target language is ਖੁਸ਼ੀ (*khushī)*. Similarly, consider the Hindi sentence प्रकाश सिंह हमारे यहाँ पधारे।. *(prakāsh siṃh hamārē yahāṃ padhārē)*. Here, प्रकाश (*prakāsh)* word is acting as proper noun and it must be transliterated and not translated because सिंह (*siṃh)* is surname and word previous to it is proper noun.

Thus, a small module has been developed for locating such proper nouns to consider them as title or surname. There is one special character '◦' in Devanagari script to mark the symbols like डा◦, प्रो◦. If this module found this symbol to be title or surname, the word next and previous to this token as the case may be for title or surname respectively, will be transliterated not translated. The title and surname database consists of 14 and 654 entries respectively. These databases can be extended at any time to allow new titles and surnames to be added. This module was tested on a large Hindi corpus and showed that about 2-5 % text of the input text depending upon its domain is proper noun. Thus, this module plays an important role in translation.

#### 2.6.2 Hindi Morphological analyzer

This module finds the root word for the token and its morphological features.Morphological analyzer developed by IIT-H has been ported for Windows platform for making it usable for this system. (Goyal V. and Lehal G.S.,2008a)

#### 2.6.3 Word-to-Word translation using lexicon lookup

If token is not a title or a surname, it is looked up in the HPDictionary database containing Hindi to Punjabi direct word to word translation. If it is found, it is used for translation. If no entry is found in HPDictionary database, it is sent to next sub phase for processing. The HPDictionary database consists of 54,127 entries.This database can be extended at any time to allow new entries in the dictionary to be added.

#### 2.6.4 Resolving Ambiguity

Among number of approaches for disambiguation, the most appropriate approach to determine the correct meaning of a Hindi word in a particular usage for our Machine Translation system is to examine its context using N-gram approach. After analyzing the past experiences of various authors, we have chosen the value of n to be 3 and 2 i.e. trigram and bigram approaches respectively for our system. Trigrams are further categorized into three different types. First category of trigram consists of context one word previous to and one word next to the ambiguous word. Second category of trigram consists of context of two adjacent previous words to the ambiguous word. Third category of the trigram consists of context of two adjacent next words to the ambiguous word. Bigrams are also categorized into two categories. First category of the bigrams consists of context of one previous word to ambiguous word and second category of the bigrams consists of one context word next to ambiguous word. For this purpose, the Hindi corpus consisting of about 2 million words was collected from different sources like online newspaper daily news, blogs, Prem Chand stories, Yashwant jain stories, articles etc. The most common list of ambiguous words was found. We have found a list of 75 ambiguous words out of which the most frequent are से *sē* and और *aur*. (Goyal V. and Lehal G.S., 2011)

### 2.6.5 Handling Unknown Words

#### 2.6.5.1 Word Inflectional Analysis and generation

In linguistics, a suffix (also sometimes called a *postfix* or *ending*) is an affix which is placed after the stem of a word. Common examples are case endings, which indicate the grammatical case of nouns or adjectives, and verb endings. Hindi is a (relatively) free word-order and highly inflectional language. Because of same origin, both languages have very similar structure and grammar. The difference is only in words and in pronunciation e.g. in Hindi it is लड़का and in Punjabi the word for boy is ਮੁੰਡਾ and even sometimes that is also not there like घर (*ghar)* and ਘਰ (*ghar)*. The inflection forms of both these words in Hindi and Punjabi are also similar. In this activity, inflectional analysis without using morphology has been performed for all those tokens that are not processed by morphological analysis module. Thus, for performing inflectional analysis, rule based approach has been followed. When the token is passed to this sub phase for inflectional analysis, If any pattern of the regular expression (inflection rule) matches with this token, that rule is applied on the token and its equivalent translation in Punjabi is generated based on the matched rule(s). There is also a check on the generated word for its correctness. We are using correct Punjabi words database for testing the correctness of the generated word.

#### 2.6.5.2 Transliteration

This module is beneficial for handling out-of-vocabulary words. For example the word विशाल (*vishāl)* is transliterated as ਵਿਸ਼ਾਲ (*vishāl)* whereas translated as ਵੱਡਾ. There must be some method in every Machine Translation system for words like technical terms and proper names of persons, places, objects etc. that cannot be found in translation resources such as Hindi-Punjabi bilingual dictionary, surnames database, titles database etc and transliteration is an obvious choice for such words. (Goyal V. and Lehal G.S., 2009a).

### 2.7 Post-Processing

#### 2.7.1 Agreement Corrections

In spite of the great similarity between Hindi and Punjabi, there are still a number of important agreement divergences in gender and number. The output generated by the translation engine phase becomes the input for post-processing phase. This phase will correct the agreement errors based on the rules implemented in the form of regular expressions. (Goyal V. and Lehal G.S., 2011)

## 3 Evaluation and Results

The evaluation document set consisted of documents from various online newspapers news, articles, blogs, biographies etc. This test bed consisted of 35500 words and was translated using our Machine Translation system.

### 3.1 Test Document

For our Machine Translation system evaluation, we have used benchmark sampling method for selecting the set of sentences. Input sentences are selected from randomly selected news (sports, politics, world, regional, entertainment, travel etc.), articles (published by various writers, philosophers etc.), literature (stories by Prem Chand, Yashwant jain etc.), Official language for office letters (The Language Officially used on the files in Government offices) and blogs (Posted by general public in forums etc.). Care has been taken to ensure that sentences use a variety of constructs. All possible constructs including simple as well as complex ones are incorporated in the set. The sentence set also contains all types of sentences such as declarative, interrogative, imperative and exclamatory. Sentence length is not restricted although care has been taken that single sentences do not become too long. Following table shows the test data set:

**Table 1: Test data set for the evaluation of Hindi to Punjabi Machine Translation System**

| | Daily News | Articles | Official Language Quotes | Blog | Literature |
|---|---|---|---|---|---|
| **Total Documents** | 100 | 50 | 01 | 50 | 20 |
| **Total Sentences** | 10,000 | 3,500 | 8,595 | 3,300 | 10,045 |
| **Total Words** | 93,400 | 21,674 | 36,431 | 15,650 | 95,580 |

### 3.2 Experiments

It is also important to choose appropriate evaluators for our experiments. Thus, depending upon the requirements and need of the above mentioned tests, 50 People of different professions were selected for performing experiments. 20 Persons were from villages that only knew Punjabi and did not know Hindi and 30 persons were from different professions having knowledge of both Hindi and Punjabi. 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 was also calculated separately by counting the number of sentences.

#### 3.2.1 Intelligibility Evaluation

The evaluators do not have any clue about the source language i.e. Hindi. They judge each sentence (in target language i.e. Punjabi) on the basis of its comprehensibility. The target user is a layman who is interested only in the comprehensibility of translations. Intelligibility is effected by grammatical errors, mis-translations, and un-translated words.

##### 3.2.1.1 Results

The response by the evaluators were analysed and following are the results:

- 70.3 % sentences got the score 3 i.e. they were perfectly clear and intelligible.
- 25.1 % sentences got the score 2 i.e. they were generally clear and intelligible.
- 3.5 % sentences got the score 1 i.e. they were hard to understand.
- 1.1 % sentences got the score 0 i.e. they were not understandable.

So we can say that about 95.40 % sentences are intelligible. These sentences are those which have score 2 or above. Thus, we can say that the direct approach can translate Hindi text to Punjabi Text with a consideably good accuracy.

#### 3.2.2 Accuracy Evaluation / Fidelity Measure

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.

##### 3.2.2.1 Results

Initially Null Hypothesis is assumed i.e. the system's performance is NULL. The author assumes that system is dumb and does not produce any valuable output. By the intelligibility of the analysis and Accuracy analysis, it has been proved wrong.

The accuracy percentage for the system is found out to be 87.60%

Further investigations reveal that out of 13.40%:

- 80.6 % sentences achieve a match between 50 to 99%
- 17.2 % of remaining sentences were marked with less than 50% match against the correct sentences.

- Only 2.2 % sentences are those which are found unfaithful.

A match of lower 50% does not mean that the sentences are not usable. After some post editing, they can fit properly in the translated text. (Goyal, V., Lehal, G.S., 2009b)

### 3.2.2 BLEU Score:

As there is no Hindi –Parallel Corpus was available, thus for testing the system automatically, we generated Hindi-Parallel Corpus of about 10K Sentences. The BLEU score comes out to be 0.7801.

## 5 Conclusion

In this paper, a hybrid translation approach for translating the text from Hindi to Punjabi has been presented. The proposed architecture has shown extremely good results and if found to be appropriate for MT systems between closely related language pairs.

## Copyright



## Acknowlegement

We are thankful to Dr. Amba Kulkarni, University of Hyderabad for her support in providing technical assistance for developing this system.

## References

Bharati, Akshar, Chaitanya, Vineet, Kulkarni, Amba P., Sangal, Rajeev. 1997. Anusaaraka: Machine Translation in stages. Vivek, A Quarterly in Artificial Intelligence, Vol. 10, No. 3. ,NCST, Banglore. India, pp. 22-25.

Goyal V., Lehal G.S. 2008. Comparative Study of Hindi and Punjabi Language Scripts, Napalese Linguistics, Journal of the Linguistics Society of Nepal, Volume 23, November Issue, pp 67-82.

Goyal V., Lehal, G. S. 2008a. Hindi Morphological Analyzer and Generator. In Proc.: 1st International Conference on Emerging Trends in Engineering and Technology, Nagpur, G.H.Raisoni College of Engineering, Nagpur, July16-19, 2008, pp. 1156-1159, IEEE Computer Society Press, California, USA.

Goyal V., Lehal G.S. 2009. Advances in Machine Translation Systems, Language In India, Volume 9, November Issue, pp. 138-150.

Goyal V., Lehal G.S. 2009a. A Machine Transliteration System for Machine Translation System: An Application on Hindi-Punjabi Language Pair. Atti Della Fondazione Giorgio Ronchi (Italy), Volume LXIV, No. 1, pp. 27-35.

Goyal V., Lehal G.S. 2009b. Evaluation of Hindi to Punjabi Machine Translation System. International Journal of Computer Science Issues, France, Vol. 4, No. 1, pp. 36-39.

Goyal V., Lehal G.S. 2010. Automatic Spelling Standardization for Hindi Text. In : 1st International Conference on Computer & Communication Technology, Moti Lal Nehru National Institute of technology, Allhabad, Sepetember 17-19, 2010, pp. 764-767, IEEE Computer Society Press, California.

Goyal V., Lehal G.S. 2011. N-Grams Based Word Sense Disambiguation: A Case Study of Hindi to Punjabi Machine Translation System. International Journal of Translation. (Accepted, In Print).

Goyal V., Lehal G.S. 2011a. Hindi to Punjabi Machine Translation System. In Proc.: International Conference for Information Systems for Indian Languages, Department of Computer Science, Punjabi University, Patiala, March 9-11, 2011, pp. 236-241, Springer CCIS 139, Germany.

Sharma R., Goyal V. 2011b. Named Entity Recognition Systems for Hindi using CRF Approach. In Proc.: International Conference for Information Systems for Indian Languages, Department of Computer Science, Punjabi University, Patiala, March 9-11, 2011, pp. 31-35, Springer CCIS 139, Germany.