Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/98374, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Approach to Paraphrase Hindi Sentences using Natural Language Processing

Nandini Sethi, Prateek Agrawal*, Vishu Madaan and Sanjay Kumar Singh

School of Computer Science Engineering, Lovely Professional University, Phagwara – 144411, Punjab, India;
nandinisethi2104@gmail.com, prateek061186@gmail.com, vishumadaan123@gmail.com,
sanjayksingh.012@gmail.com

Abstract

Background/Objectives: Nowadays, most of the automatic work has been done in English language but not much work has been done in Hindi language. The main objective is to convert the existing sentence in different form by remaining the semantic or meaning same. This will helpful in converting the complex sentence into simpler one.Methods/ Statistical Analysis:Our system mainly deals with Hindi Sentences and its different forms. It takes a sentence as input and produces another sentence without changing its semantic after applying synonyms and antonyms replacement method.Findings:Reframing of Hindi sentences can be used to change a complex sentence in simplified form.The system is implemented in java. WampServer is used as database for our system. In our paper, we have described complete algorithm to paraphrase the sentences. The performance of the system is totally dependent on the size of Database. Application/ Improvements:This application can be helpful in designing robots to understand different forms of Hindi sentences, to use as Hindi tutor for students to get them idea about different form of sentences and in plagiarism tools to find the higher level of plagiarized text up to certain extent. The test results have been verified by various tutors.

Keywords: Antonyms, Paraphrase, Parsing, Synonyms, Transliteration

1. Introduction

Hindi, the major portion of language in India, is still in its childhood stage concerning to natural language processing research and applications. Paraphrasing or reframing of Hindi sentences is an important issue that needs to be noticed. Nowadays, there is lots of automatic work done for English language text, articles but not much automatic work has been done for Hindi language. English language categorizes the alphabets into two parts:

- Consonants: The number of consonants in English language are 21(e.g. A, F, K).
- Vowels: The number of vowels in English language are 5 i.e. a, e, i, o, u.

But Hindi language is much complex than English because Hindi language categories the alphabets into two parts:

Consonants: The numbers of consonants are 40. These consonants are known as "Vyanjan/व्यंजन" in Hindi language.

- Vowels: The numbers of vowels are 10, which include various types of symbols (ten vowels have sign (matras), half letters & halants etc.).

So, due to these differences between both the languages the methods which are applicable for English language are not able to be directly used for the systems of Hindi language.

Different ways to paraphrasing a sentence:

1. Replacing words with its synonyms.
2. Replacing words with equivalent antonyms.
3. Rearranging the order of sentences on its priority etc

Various activities involved in the process of paraphrasing:

*Author for correspondence

1.1 Paragraph Segmentation

Firstly the text is divided into segments or different sentences. This division of sentences is based on the presence of some symbols in Hindi language that indicate the completion of the sentence. In Hindi language "|", this symbol is considered as the completion of sentence. Some other symbols like "?,;" also represent the completion of sentences. So on the basis of these symbols our system will divide the whole paragraph into different sentences. These sentences are used for further processing.

1.2 Syntax Analysis (Parsing)

The main scope of this process is parsing. It plays an essential role in understanding many language systems. In this step, a simple Hindi sentence is granted as an input and then which is converted into a hierarchical form which will address to the units of meaning in the sentence[1]. It uses the first components of the token to produce tree probably intermediate structure that depicts the grammatical structure of the token stream.

There are distinct parsing formalisms and algorithms in which formalism has two leadingcomponents:

1. Grammar: A declaratory representation that portray the syntactic structure of sentences in the language.
2. Parser Based on the morph-syntax rules, it captivate the input and gives output as a syntax tree.

1.3 Semantic Analysis

Semantic analysis is used to check whether inserted sentence is accurate or not. Although the main intend of semantic analysis is the formation of the target language representation of the sentences meaning which indicate assigning meanings to the structures created by syntactic analysis. Semantic can play an important role in option among contend in syntactic analyses.

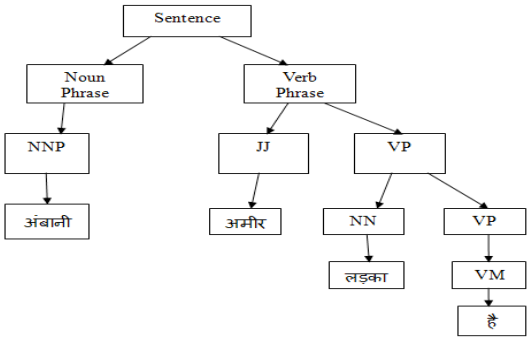

Figure 1. The result of syntactic analysis of "अंबानीअमीरलड़काहै".

1.4 Reframing Rules

In the database the knowledge is represented in the forms of rules and facts. These facts and rules are applied to get the desired output. The rules can be certain patterns or symbols that matches with the input and for applying the processing.

NLP is the scientific study of natural languages and a field of computer science which makes computer interact-able with the human beings. NLP is used to train the computer for various natural languages. There are many challenges for interaction between computers and humans1. Computers understand only binary digits but humans can't deal with binary digits. So we require huge database stored in our system for processing of human understandable words by computers. NLP is one of the most promising technique through which humans can interact with computers.

Since, NLP provides a technique through which human can interact with computer in any natural language or NLP is a field which is used to making computer understandable about various human languages. But making computer understandable about human language is a very difficult task2. First and the foremost challenge is the platform or the framework through which humans can interact with computer. NLP technique can be used to reframe Hindi sentences using various techniques. Reframing converts the Hindi sentence into another sentence without changing its meaning. One of the best suitable languages for working of NLP is Java.

In the existing paraphrasing system such as Ginger software3, developers have used a method using synonym, idioms and phrases replacement for English language. Marcel Bollmann (2011)performed syntax analysis in German language and they focus on the order of the words4 for its implementation. Use a standard (nonparallel) monolingual corpus to generate paraphrases, based on dependency graphs and distributional similarity for English language5presented a system for compositional machine transliteration, in which multiple transliteration components can be composed. Their system can enable transliteration functionality between the languages even when no direct parallel names corpus exists between them. They demonstrate the performance and functionality

benefits of the compositional methodology using a state of the art of machine transliteration framework in English and a set of Indian languages, like, Kannada, Hindi, and Marathi6. N-Gram technique was used7 to identify the similar texts in two documents written in Hindi. Stemming of words to identify the root words was done in Hindi language8. Also most of the existing systems are static, only one form of new paragraph is given as output and are made for text in English language while in the proposed system is for text in hindi language and there are different output screens for antonyms, synonymsand one overall result including the these three things.

2. Our Approach

Proposed system is done in various steps like text segmentation, tokenization, parts of speech tagging, classification of sentences, database generation, reframing rules etc shown in Figure 2.

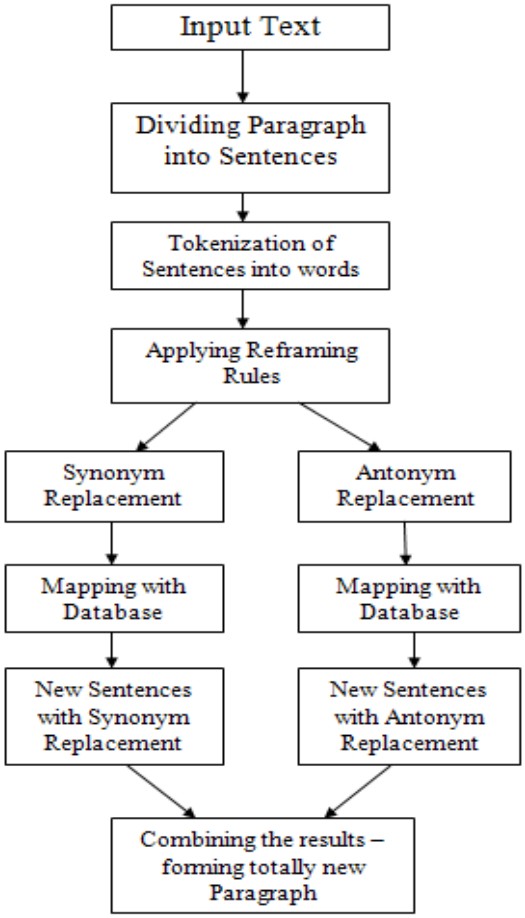

Figure 2. Processing steps of proposed approach.

2.1 Input Text

In this system the user can input the text in 3 ways:

1. User can input the text in Hindi by using Hindi keyboard provided by our interface. The user need to the click on the button of Hindi keyboard and type with the help of that keyboard9. This type of keyboard will be helpful to those who are not aware of using English keyword to type Hindi alphabets. This can also be use to teach students about various matras and word formation in an innovative and interesting way. User can easily display the keyboard by clicking on the button of virtual keyboard and also easily exit from this keyboard by pressing the exit button on keyboard. There is also an option available for reset the whole text which will clear the existing text in the input box. An important feature of this keyboard is that the alphabets are arranged in this keyboard are in such an order in which teachers teach them. This order makes the keyboard user friendly because user can easily find the alphabets due to this sequence. The interface having keyboard is shown in Figure 3. Then, user can exit from this keyboard section by just clicking on the exit button present on the keyboard.
2. Secondly, user can input the text by the process of transliteration. Transliteration is the process which will automatically convert the text typed in English to Hindi language. This process will help the user to easily enter the input10 For this algorithm we will make use of Hindi language Unicode's which will make the mapping between English and Hindi alphabets and will generate the desired output i.e. text in Hindi language.
3. Third approach is by simply pasting the text in the input text area in Hindi language.

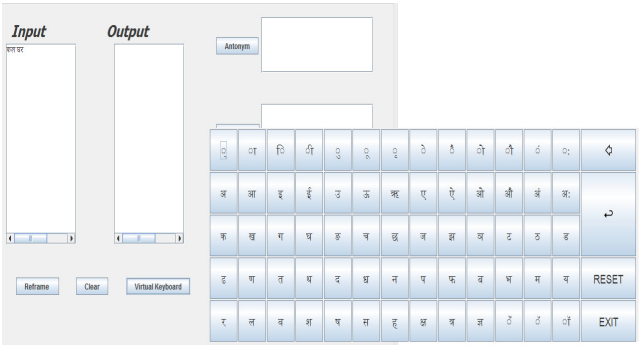

Figure 3. Hindi keyboard.

2.2 DividingText or Paragraph into Sentences

Firstly the text is divided into segments or different sentences. This division of sentences is based on the presence of some symbols in Hindi language that indicate the completion of the sentence. In Hindi language "|", this symbol is considered as the completion of sentence. Some other symbols like "?,;" also represent the completion of sentences. So on the basis of these symbols our system will divide the whole paragraph into different sentences. These sentences are used for further processing.

2.3 Tokenization of Sentences into Words

Various grammatical rules are applied for tokenization of sentences into words. This step is completed with the help of tagger which is used by our system for tagging the words like nouns, pronouns, adjectives etc.This tagger contains two types of files: first is the file containing Hindi words with their tags like a dictionary file and second file contains various grammar rules which specify which word will act as noun or adjective at which place of sentence.

2.4 Applying Reframing Rules

Reframing rules are those which when applied to the sentences change the syntax of the sentence but the meaning of the sentence remains the same. These rules are present in the database for these replacements and for the generation of new sentences or paragraph.

2.4.1 Synonyms Replacement

Two or more interrelated words that can be changed in a context are synonyms11,12. In this system will match the word with the database if the word found in the database then that word will be replaced with its synonym and the system will search for the next word from the entered text to be mapped with database and this process will continue till the last word of the entered text. Some of the sample synonym replacements of words are shown in Table 1.

Proposed Algorithm for Synonym Replacements

```
Synonym_replace(input text)
{
      Connectivity with Database;
      list_of_synonym = add all the proverbs present in
      database to this array list;
      Iterator iterator = list_of_synonym.iterator ();
      while(iterator.hasNext())
      {
keyword = rst.getString(2);
      if(input.contains(keyword))
      {
      syn= rst.getString(3);
      new_sen=input.replace(keyword, syn);
      input=new_sen;
      }
   }
}
```

2.4.2 Antonym Replacement

When a word expresses the meaning opposite to the given word, it is referred as antonym[13,14].In this proposed system negative sentences can be converted to positive sentences by using specific antonym. This can done vice versa also i.e. the positive sentences are converted into negative sentences also.Some of the sample antonym replacements of words are shown in Table 2.

Table 1. Synonym replacement

Word	Sample synonyms
अतिथि	मेहमान, अभ्यागत, आगन्तुक, पाहूना
अमृत	सुधा, सोम, पीयुष, अमयि, अमी
अग्नि	अनल, पावक, वहनि, कृशानु, शखिी
अंधकार	तम, तिमिर, तमिस्र, अँधेरा
जल	नीर, तोय, वारि, अमृत, उदक, अंबु, पानी

Table 2. Antonym replacement

Word	Sample Replacement
विष नही	अमृत
अकाल नही	सुकाल
सम्मान नही	अपमान
खुश नही	उदास
बड़ा नही	छोटा

Proposed Algorithm for Antonym Replacement

```
Antonym_replace(input text)
{
      Connectivity with Database;
      list_of_antonym = add all the proverbs present in
      database to this array list;
      Iterator iterator = list_of_antonym.iterator ();
      while (iterator.hasNext())
      {
keyword = rst.getString(2);
      if (input.contains(keyword))
      {
   ant= rst.getString(3);
      new_sen=input.replace (keyword, ant);
      input=new_sen;
      }
   }
}
```

2.5 Combiningthe Results Forming Entirely New Paragraph

In this step, both the synonym replacement and antonym replacement are combined are it will form a purely new paragraph with same meaning but different syntax and words. This paragraph contains both the antonym replaced sentences and synonym replaced sentences. The reframing is the restatement of the paragraph without changing the meaning of the actual text. It explains the paragraph in simple words and enables to present the same paragraph in different ways.

Here, the sentence "राम का अपमान हुआ और वह आकाश की तरफ देखने लगा" is, reframed using synonym replacement rules, antonym replacement rules. Here the actual meaning of the sentence is kept same as the original sentence. There is either change in the sequence of the word or, word's synonym or antonym is used. The reframing does not accompany the direct reference; it serves as a source to new reframed paragraph.

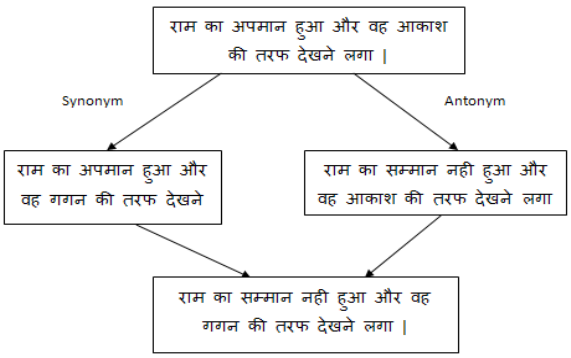

Figure 4. Reframing types.

Proposed algorithm

```
Paraphrase (sentence or paragraph)
INPUT Text (any of three methods)
Apply Text Segmentation.
Set count=1
While (count < number_of_sentences)
   {
      Synonym_replace(input);
      Antonym_replace(input);     //Calling Of Functions
}
Count++
Return Output
End
```

3. Results

This work will generate new sentences based on specified rules and after synonyms and antonyms replacements. Figure 5. shows the main interface of our proposed system and Table 3 shows the Result analysis of the system.

In this Result analysis, it shows various types of user input, it can be affirmative sentences or negative sentences and their corresponding results generated by the system. The results are shown in three different windows. One is for antonym replacement; another is for synonym replacement and third is for showing the combined result.

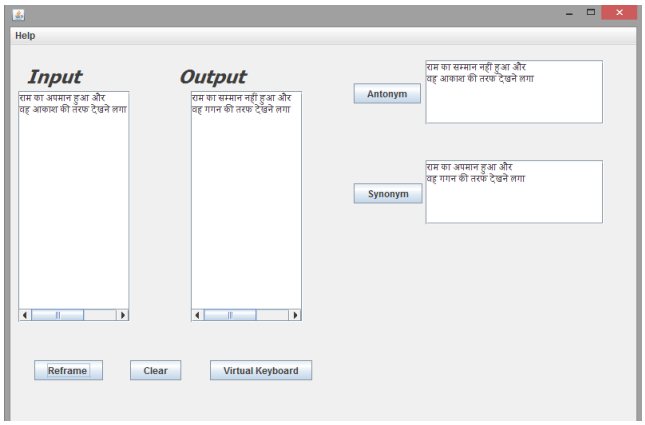

Figure 5. Interface of the system

Table 3. Result analysis

Input	Replacement and Reframing	Result
राम का अपमान हुआ और वह आकाश की तरफ देखने लगा	Synonym	राम का अपमान हुआ और वह गगन की तरफ देखने लगा
	Antonym	राम का सम्मान नहीं हुआ और वह आकाश की तरफ देखने लगा
	Synonym +Antonym	राम का सम्मान नहीं हुआ और वह गगन की तरफ देखने लगा
मोहन ने अमृत पयिा	Synonym	मोहन ने पीयुष पयिा
	Antonym	मोहन ने वषि नहीं पयिा
	Synonym +Antonym	मोहन ने वषि नहीं पयिा
सोहन के घर धन का अकाल नही आया	Antonym	सोहन के घर धन का सुकाल आया |
	Synonym	सोहन के घर दौलत का अकाल नही आया |
	Antonym +Synonym	सोहन के घर दौलत का सुकाल आया |

4. Conclusion

In this proposed system we have discussed sentence reframing technique using NLP. This system can be used by scholars, technical writers and researchers. This system can be further extended to develop software for semantic analysis for information extraction and other. This system can be helpful in making a robot understand different forms of sentences. It can also be used to develop an intelligent system that can take decisions like humans. This work can be extended to make a decision support system that will work is a similar manner like humans and can respond just like humans by understanding the different forms of sentences given to it as an input.

5. References

1. Patterson DW. Introduction to AI andexpert systems. Prentice Hall; 1990.
2. Knight R. Artificial intelligence. 3rd edition. Tata McGraw Hill; 2009.
3. Ginger Software [Internet].[Cited 2016 Jan 15]. Available from: http://www.gingersoftware.com/products/sentence-rephraser.
4. Bollman B, Marcel M. Adapting Simplenlg to German. Proceedings of the 13th European Workshop on Natural Language Generation (ENLG-11); 2011.p. 133–8
5. Lin D, Pantel P. DIRT – Discovery of Inference Rules from Text.Proceedings ofACM SIGKDD Conference on Knowledge Discoveryand Data Mining;2001.p. 323–8.
6. KumaranA, KhapraM, Bhattacharyya P.Compositional machine transliteration. ACM Transactions on Asian Language Information Processing.2010; 9(4):1–29.
7. Urvashi G,Maulik GV. Maulik: A plagiarism detection tool for Hindi documents. Indian Journal of Science and Technology.2016 Mar; 9(12):1–11.DOI: 10.17485/ijst/2016/v9i12/86631.
8. Leena J, Prateek A.Text independent root word identification in Hindi language using natural language processing. International Journal of Advanced Intelligence Paradigms. 2015;7(3/4):240–9.
9. Keyboard [Internet]. [Cited 2016 Feb 19]. Available from: http://www.baraha.com/help/Keyboards/dev_phonetic.html.
10. Haque R, Dandapat S, SrivastavaAK, NaskarSK. English-Hindi transliteration using context-informed PB-SMT: the DCU System for NEWS 2009.NEWS '09 Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration; 2009.p. 104–7.
11. Synonym List[Internet]. [Cited 2016 Jan 01]. Available from: http://hindilearner.com/hindi_words_phrases/hindi_synonyms.html.
12. Synonym List[Internet]. [Cited 2016 Jan 01]. Available from: http://www.hindigrammaronline.com/2013/02/hindi-synonyms.html.
13. Antonym List[Internet]. [Cited 2016 Jan 21]. Available from: http://www.englishkitab.com/Vocabulary/Antonyms/Antonyms-2.html.
14. Antonym List[Internet]. [Cited 2016 Jan 21]. Available from: http://www.hindikunj.com/2009/11/antonyms-in-hindi.html.