Hybrid Approach for English to Punjabi Translation System for News Paper Headlines in a Specific Domain

Savita Singla
M.Tech, CSE
Yadavindra College of Engineering
Talwandi Sabo, India

Prof. Seema Baghla
Assistant Professor, CSE Department
Yadavindra College of Engineering
Talwandi Sabo, India

Abstract

Abstract – Machine translation (MT) plays an important role in benefiting linguists, sociologists, computer scientists, etc. by processing natural language and to translate it into some other natural language. The demand of translation has become more in recent years due to increase in the exchange of information between various regions across the world. Due to this reason machine translation has become an important research subfield and comes under the Artificial Intelligence(AI). Many approaches have been used in the recent times to develop an MT system. Each of these approaches has its own advantages and disadvantages.

The performance of an MT system depends on the approach used to design the system. In this paper we are presenting a brief overview of the MT and various techniques of designing an MT system. Also we are discussing the challenges faced while translating one language into another.

Index Terms- Machine Translation, Rule Based Approach Direct Approach, Transfer Based Approach, Inter lingual approach, Hybrid Approach, EBMT, SMT.

1. Introduction

Machine translation is very important field of natural language processing. It can be defined as the study of designing the systems that can translate one human language into another. These systems take input in one natural language and convert it into another human language with the help of various machine translation techniques. Rule based machine translation technique , example based machine translation technique and statistical machine translation techniques are most commonly used for translation. In this paper we use hybrid approach to translate newspaper headlines of English Language into Punjabi language equivalent. The language that is given as an input is called Source Language (English Text) and the language in which we get the output is called Target language(Punjabi Text).

2. Need for Translation

India is a multilingual country but most of the Indian states population is not familiar with English. Information is available on world wide web and on various another resources is in English. And the people who doesn't understand English properly can't make use of this wide information. In India there are more than 50 recognized languages in the world and Punjabi is very important among these Indian languages. In is widely spoken and use for official work by the government of Punjab. Most of the people in Punjab are familiar with Punjabi but not with English because most of the population in Punjab is living in villages and in small towns. This system makes these people to understand the headlines of English newspapers. Proposed system is made as an online system so that it can be used anywhere with the help of internet. This online interface for English to Punjabi Translation helps the people in Punjab to read English news paper headlines anywhere where internet facility is available.

3. Literature survey

Kamal Deep, Goyal V. (2011) The system developed by Kamal Deep & Dr. Vishal Goyal named **Punjabi to English Transliteration System** using a rule based approach and achieved accuracy of 93.23%. Transliteration scheme uses grapheme based method to model the transliteration problem. This system addresses the problem of forward transliteration of person names from Punjabi to English by set of character mapping rules. This system is accurate for the Punjabi words but not for the foreign words. System evaluated for names from the different domains like Person names, City names, State names, River names, etc. [1]

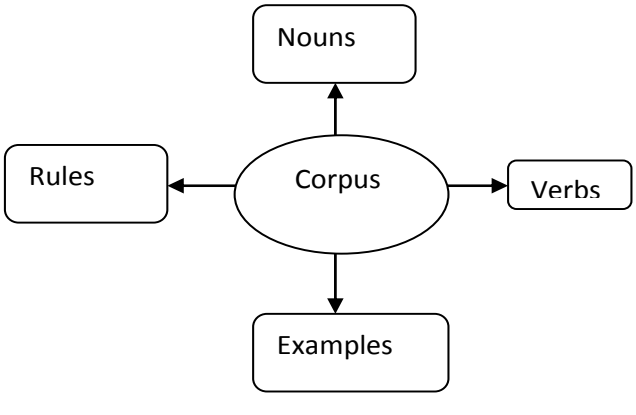
R.M.K. Sinha and A. Jain, AnglaHindi: An English to Hindi Machine-Aided Translation System. presented a machine translation system called AnglaHindi which is an English to Hindi version of the ANGLABHARTI translation methodology with a mixture of some example-based translation methodology. AnglaHindi system has been web enabled and is available at URL: http://anglahindi.iitk.ac.in for free translation. The system generates approximately 90% acceptable translation in case of simple, compound and complex sentences up to a length of 20 words. [2]

K.K. Batra , G.S. Lehal , Automatic Translation System from Punjabi toEnglish for Simple Sentences in Legal Domain

The system has been developed to translate simple sentences in legal domain from Punjabi to English. Since the structure of both the languages is different, direct approach of translating word by word is not possible. So, indirect approach i.e. rule based approach of translation is used. The system has analysis, translation and synthesis component. The steps involved are pre processing, tagging, ambiguity resolution, phrase chunking, translation and synthesis of words in target language. The accuracy is calculated for different phases of the system and the overall accuracy of the system for a particular type of sentences is about 60%.[3]

4. Research Methodology

In the proposed system hybrid approach is used to translate the news headline written in English text into its equivalent Punjabi text. Hybrid approach in the proposed system is a combination of Rule based approach; Example based approach and direct mapping approach. Proposed system uses a corpus for translation purpose. Corpus consist of rules , examples , nouns , verbs and other entities stored within database and within the programming logic.



Direct Mapping approach for translation is used for such sentences for which translation using dictionary is not possible. For example in case of idioms direct translation from source to target language is not possible and hence in that case direct mapping is to be used.

For Eg: nothing to hide

ਲੁਕਾਉਣ ਦੇ ਲਈ ਕੁਝ ਵੀ ਨਹੀ

Approach of direct mapping translate the sentence in two phases :

1. Pre processing :- in this phase if source sentence is a combination of two or more sentences joined with ","(Comma) , ":" (Semi Colon) or with another special symbol then divide the sentence in multiple independent sentences and parse them separately to obtain the final result.
2. Direct Translation : In this phase direct mapping is done from the corpus.

Example based approach for translation is used for similar types of sentences. For example there are such sentences in which there are many common words, in such cases example based approach for translation is better approach.

For Eg: Child dies in mishap

ਦੁਰਘਟਨਾ ਵਿੱਚ ਬੱਚੇ ਦੀ ਮੋਤ

Example based approach translate the sentence in two phases which are as follows :

1. Pre Processing : - in pre processing source sentence is parsed to find the match with the example stored in corpus.
2. Identification of Nouns : After the sentence is parsed and example is found, nouns from the source sentence are to be found from the corpus. These nouns can also be combination of more then one nouns and have to join to form a single entity.
3. Translation Phase : in this phase final step of translation is done with the help of existing translated example from the corpus.

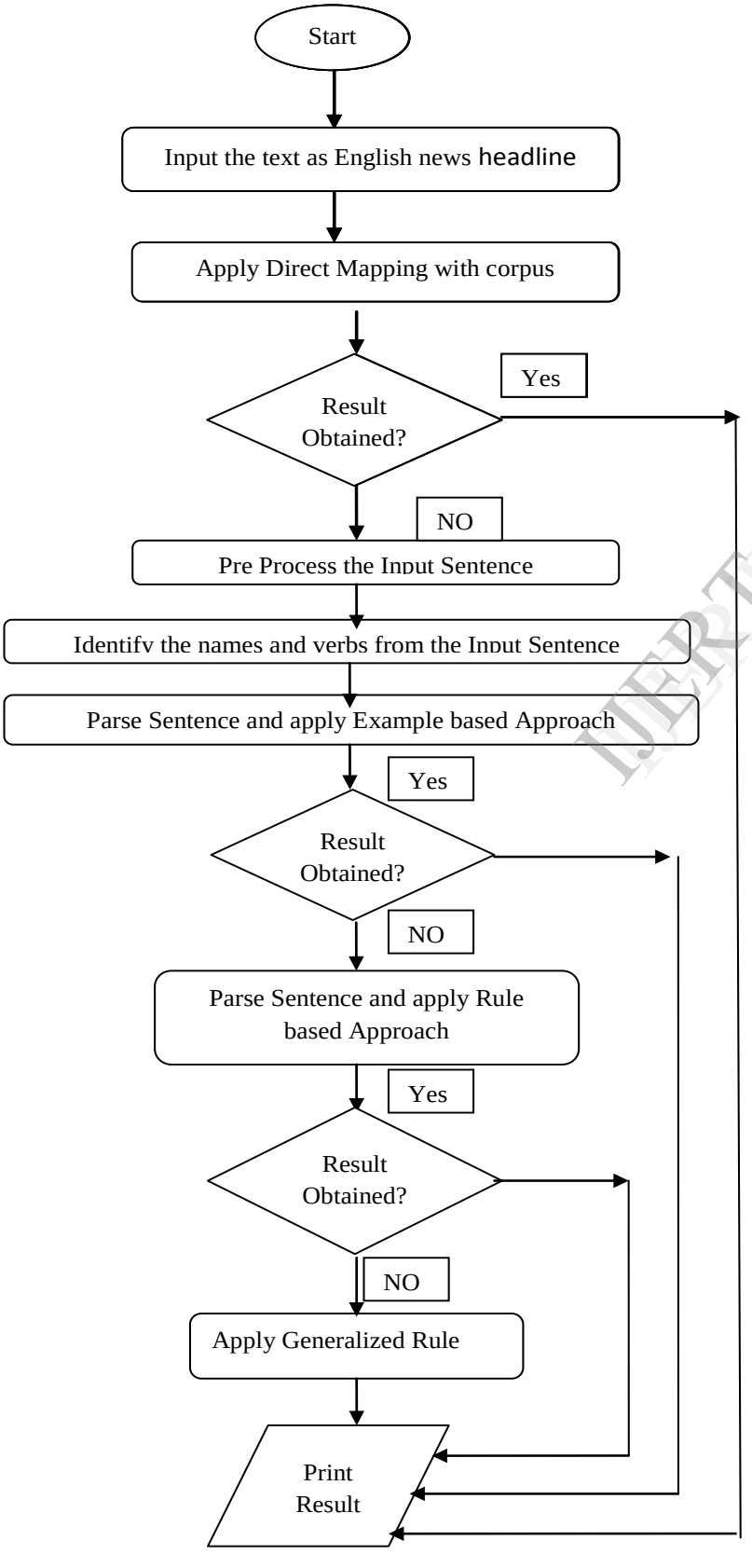
Rule based approach in the proposed system is used for such sentences in which grammatical rules of both source and target languages can be applied.
Rules are made using programming language by taking consideration the features of both source and target language.
Rule based approach translate the source sentence into target sentence in the following phases :

1. Pre Processing : - In this phase tokenization of words is to be done by extracting them from the sentence. These words are to stored in the independent memory locations.
2. Identification of Nouns and Verbs :- From the extracted words nouns and verbs are to be indentify and assign the corresponding mark to them.
3. Translation Phase : In this phase various rules which have been created are applied

to these extracted words and final target sentence is generated.

5. Flow Chart

The method for translation is displayed in the following flow chart :



As shown in the above flowchart, proposed system first check if there is a direct sentence corresponding to the source sentence and if there exist such a sentence then result is printed otherwise sentence is parsed and example based approach is applied on the sentence and again system check for the result and if result is obtained then it is printed otherwise rule based approach for translation is applied on source sentence and final result is printed.

6. System Results

The proposed system is checked on news paper headlines of various newspapers under various sections like sports , education , political , regional etc. And system is giving very good results. Some of the statistics about the results are as in the following table:

Set	No. Of Examples
Data base Entries for nouns	15,000
Testing	300
Overall Accuracy	84%

7. Conclusion and Future scope

In this paper we present the technique for translation the news paper headlines of English language into its equivalent Punjabi language. As shown our proposed system is generating very good results but these results can also be improved by using statistical machine translation approach in which system can translate the input sentence by using existing translated sentences. In addition to this approach database can be improved by storing more names, dictionary words etc. The system is currently working on news paper headlines and can be extended to any type of sentences in future. Statistical Approach for machine translation can also be to improve the results and to make it platform independent.

8. References

[1] K. Deep, Dr. V. Goyal "Hybrid Approach for Punjabi to English Transliteration System," International Journal of Computer Applications (0975 – 8887) Volume 28– No.1, August 2011.

[2] R.M.K. Sinha and Ajay Jain, AnglaHindi: An English to Hindi Machine Translation System, MT Summit IX, New Orleans, USA, Sept.23-27, 2003.

[3] K.K. Batra , G.S. Lehal , Automatic Translation System from Punjabi toEnglish for Simple Sentences in Legal Domain

[4] Latha R. Nair and David Peter S., Machine Translation systems for Indian Languages, IJCA, Feb 2012.

[5] Cheragui M.A., "Theoretical Overview of Machine Translation", Proceedings ICWIT 2012.

[6] Kamaljeet Kaur Batra, and G S Lehal ,Rule Based Machine Translation of Noun Phrases from Punjabi to English.

[7] R.M.K. Sinha and Anil Thakur, Divergence Patterns in Machine Translation between Hindi and English, 10th Machine Translation summit (MT Summit X), Phuket,Thailand, September 13-15, (2005), 346-353.

[8] Aniket Dalal, Kumara Nagaraj, Uma Sawant,Sandeep Shelke and Pushpak Bhattacharyya, Building Feature Rich POS Tagger for Morphologically Rich Languages, ICON 2007, Hyderabad, India, Jan, 2007.

[9] Akshar Bharati, Vineet Chaitanya, Amba P. Kulkarni,Rajeev Sangal Anusaaraka: Overcoming the Language Barrier in India. (informal publication) Electronic Edition (link) BibTeX [cs.CL/0308018]

[10] Computational Paninian Grammar for Dependency Parsing Dipti Misra Sharma,LTRC, IIIT,Hyderabad, NLP Winter School 25-12-2008

[11] Akshar Bharati, Rajeev Sangal: Parsing Free Word Order Languages in the Paninian Framework. ACL 1993

[12] Akshar Bharati, Rajeev Sangal: A Karaka BasedApproach to Parsing of Indian Languages. COLING 1990: 25-29

[13] R M K Sinha, Some thoughts on computer processing of natural Hindi.. Annual convention of Computer Society of India, 1978, pp 151-165.