Web Based Hindi to Punjabi Machine Translation System

Vishal Goyal and Gurpreet Singh Lehal
Department of Computer Science, Punjabi University, Patiala, Punjab, India
{vishal.pup,gslehal}@gmail.com

***Abstract* - Hindi and Punjabi are closely related languages with lots of similarities in syntax and vocabulary Both Punjabi and Hindi languages have originated from Sanskrit which is one of the oldest language. In terms of speakers, Hindi is third most widely spoken language and Punjabi is twelfth most widely spoken language. Punjabi language is mostly used in the Northern India and in some areas of Pakistan as well as in UK, Canada and USA. Hindi is the national language of India and is spoken and used by the people all over the country. In the present research, Basic Hindi to Punjabi machine translation system using direct translation approach has been developed. The results of this translation system are surprisingly good. The system includes lexicon based translation, transliteration and continuously improving the system through machine learning module. It also takes care of basic word sense disambiguation.**

***Index Terms* - Machine Translation System, Closely related Languages, Hindi, Punjabi, Natural Language Processing, Computational Linguistics, Transliteration.**

I. Introduction

MT is a field of research that has been around since the birth of electronic computers. Warren Weaver, a director of the Rockefeller Foundation received much credit for bringing the concept of MT to the public when he published an influential paper on using computers for translation in 1949. MT is the name for computerized methods that automate all or part of the process of translating from one human language to another. Fully-automatic general purpose high quality machine translation system (FGH-MT) is extremely difficult to build. In fact, there is no system in the world of any pair of languages which qualifies to be called FGH-MT. This paper explains the methodology followed for developing the machine translation system between closely related languages – Hindi and Punjabi. Closely related languages have lots of similarities in syntax and vocabulary. Machine Translation between closely related languages is easier than between language pairs that are not related with each other. Having many parts of their grammars and vocabularies in common reduces the amount of effort needed to develop a translation system between related languages. Closely related languages are the languages of people who have similar cultures and common historical roots.

II. History

The first attempt to verify the hypothesis that related languages are easier to translate started in mid 80s at Charles University in Prague [FEMTI; Hajic et al. 2000]. The project was called RUSLAN and aimed at the translation of documentation in the domain of operating systems for mainframe computers. From that date to till date so many examples are there in history which support the argument that with close languages, the quality of MT system, with simple techniques, is better. To name a few one are CESILKO (a system for translating Czech and Slovak), MT system for translating Turkish To Crimean Tatar etc. We are also trying to strengthen the same concept by experimenting with a word for word direct translation system for Hindi to Punjabi. These languages are very closely related and have many features in common.

III. System Description

The major task behind direct Machine translation system is developing an exhaustive lexicon consisting of source language words along with its corresponding translated version of the target language. There is no machine readable dictionary available for Hindi to Punjabi language. Two traditional dictionaries – one from Bhasha Vibhag, Patiala and another from National Book Trust has been published. We got it digitized and moulded required for machine translation purpose which is itself a big job. Now lexicon consists of approx. 1,00,000 words. This lexicon is used for word for word translation. Extending the lexicon demands large hindi corpus. Sometimes it is available in Unicode format and sometimes it is available in number of other non standard fonts like Susha, Agra, Krutidev etc. which needs to be converted into Unicode first. Conversion is being done through Font Converter that converts non standard fonts into Unicode Format. The beauty of the developed Font converter is that it is able to convert MS-Access files, MS Word files, HTML files and Text Files. Because the corpus can be in any file type, so it handles all the possible file types. Besides translation it performs the task of transliteration as well. Transliteration means to replace character by character of word from source language character to target language character like प्रेमचंद into ਪ੍ਰੇਮਚੰਦ. This system is basically an extension

doi:10.4304/jetwi.2.2.148-151

of previous system that does not handle any word sense disambiguation. The above said system just checks the words to be translated in the dictionary, if found it is replaced with the translated version stored in the dictionary, otherwise it is transliterated. But now in the extended system, the lexicon is divided into two parts – one table consists of words with no disambiguation and second consists of words that have multiple meanings depending upon the context of the word in which it has been used in the sentence.

IV. System Architecture

The architecture for the HPMTS (Hindi to Punjabi Machine Translation System) consists of number of modules that are listed below:

a. Training the system with training corpus
b. Input Text Font Conversion into Unicode Format
c. Hindi Text Normalization
d. Finding and Replacing Collocations
e. Finding and replacing named entities
f. Word to word translation using lexicons
g. Resolving Ambiguity among words
h. Transliteration of words
i. Post Processing
j. Improving the accuracy of the system through machine learning during every translation job.
k. Testing the system using test corpus other than train corpus

In the above architecture, the most important part and starting point is to train the system. Train the system means generating the lexicon using the already existing corpus. The second module is optional and is skipped if the inputted text is already in Unicode format. Unicode Font requirement arises due to internalization of the system and making the system free from specific font dependency. This font converter can be also used for converting the non-Unicode corpus into Unicode format corpus. Indian language words face spelling standardization issues, thereby resulting in multiple spelling variants for the same word. The main reason for this phenomenon can be attributed to the phonetic nature of Indian Languages and multiple dialects. To give an idea of this data problem, these words were found – मंजिल, मन्जिल, मंज़िल Third module is Hindi Text Normalization that solves this spelling variant problem. Hindi text is normalized into standard spellings before it goes for translation. Next Module of the system find and replaces all the collocations using the lexicon enteries. A Collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. Or in the other words of Firth (1957:181) : "Collocations of a given word are statements of the habitual or customary places of that word". This module helps in increasing the accuracy of the translation. Generating Lexicon for Collocations is itself a challenging task. Then comes the turn of the heart of the system – word for word translation uses the lexicon. This search for the Hindi word in the lexicon and replaces it with the corresponding Punjabi translated version present in the lexicon. If this Hindi word is not found in the lexicon it searches that word in the database of ambiguous words, if found using tri-gram approach it resolves the ambiguity of word and replaces it with correct Punjabi meaning among multiple Punjabi meanings. For Example, the hindi word ਸਰੂਪ can be translated into either of the two Punjabi words - ਸਮਾਨ, ਸੁੰਦਰ. But how will the system decide which word to choose is basically to know the context in which the Hindi word ਸਰੂਪ has been used in the sentence. If the word is not found in both the tables it means it is not available in the database and need to be transliterated. For improving the accuracy of the system, this is must to know the system about which new words have been come across and if they have been transliterated accurately or not. If they were not present in the database and need to be present, it is added to lexicon for future translations. If it has been translated wrongly but required one, it is corrected first before adding to the lexicon. In this way this is the ongoing improvement of the system performance during every translation exercise through machine learning module. Post Processing Module takes into consideration some common grammatical mistakes that has been done during translation phases and based on the rules framed, it removes those mistakes and increases the accuracy to the system. Now system has been trained a lot by number of translation exercises, it is time to check the accuracy of the system by testing the system through test data other than the data used for training. Testing the system is also very tedious task. First step in it is to prepare the test cases that covers all the possibilities.

V. Web Based Tool

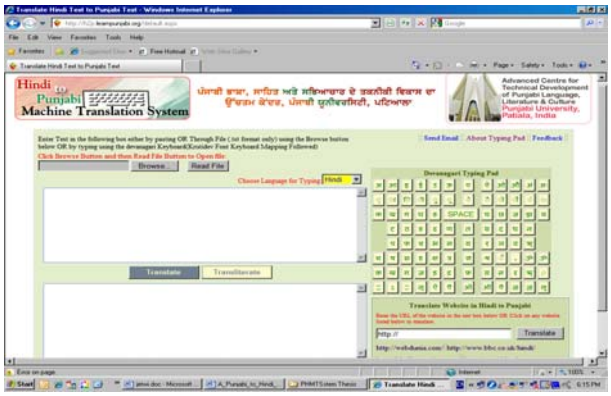
Research must not be restricted to papers, It must be propagated to public for use and test. Taking this aim, the whole system has been developed as a web tool and is online for use fre of cost. The website address is http://h2p.learnpunjabi.org/ . Following are the features of this web tool:

a. Hindi Text can be written in Unicode encoding by using the most popular Hindi Font Krutidev. This concept is very useful for those who are in habit of typing the text in Krutidev and later they find some source for converting them into Unicode encoding. Thus, this feature has solved their purpose in very easy manner. Now, they will type in their style and the typed matter will also be in Unicode.
b. The text can also be input to the system for translation through text file. File can be read using the Browse button provided.
c. Input text can be translated into Punjabi text by just clicking the Translate button. Within seconds, the text is translated into Punjabi.

d. Input Text can be transliterated also, if there is need by clicking on the Transliterate button.
e. Email facility has also been provided. Text can be typed in Hindi and English both. Subject can be written in both the languages. Then while sending email there is an option of sending the email in original or after translating in Punjabi. This is also very powerful feature.
f. The main feature of the webtool is user can translate the full Hindi website into Punjabi website by just providing the link of hindi website and press Translate. The hindi website is translated into Punjabi within seconds and Punjabi version of the website is displayed in the same format as it was originally in hindi website.

VI. Screen Shots

Following is the screen shot of web based machine translation system:



VII. Sample Output

A. Input Text:

भारत और पड़ोस

गुरुवार, **01** नवंबर, **2007** को **08:02 GMT** तक के समाचार

कर्नाटक को लेकर राजनीतिक सरगर्मी बढ़ी

कुमारस्वामी ने राजनीतिक चाल बदलते हुए येदियुरप्पा को मुख्यमंत्री के रुप में समर्थन देने का फ़ैसला किया है

कर्नाटक को लेकर राजनीतिक गहमागहमी तेज़ हो गई है और अब यह मामला दिल्ली आ गया है.

संभावना है कि जल्दी ही कर्नाटक के राजनीतिक भविष्य का कोई फ़ैसला हो जाएगा.

बुधवार को जहाँ भाजपा के उच्च स्तरीय प्रतिनिधि मंडल ने प्रधानमंत्री मनमोहन सिंह से मुलाक़ात की है प्रधानमंत्री और सोनिया गाँधी ने कांग्रेस कोरग्रुप की बैठक में कर्नाटक के राजनीतिक हालात पर चर्चा की है.

उधर भाजपा ने धमकी दी है कि यदि भाजपा-जनतादल (एस) को सरकार बनाने के लिए आमंत्रित नहीं किया गया तो दोनों दलों के **129** विधायक राष्ट्रपति के सामने परेड करेंगे.

उल्लेखनीय है कि जनता दल (एस) और भाजपा का गठबंधन सात अक्तूबर तक सत्ता में था. लेकिन समझौते के अनुसार मुख्यमंत्री बदले जाने के विवाद के चलते गठबंधन टूट गया और भाजपा ने सरकार से समर्थन वापस ले लिया.

B. Translated Output Text:

ਭਾਰਤ ਅਤੇ ਗੁਆਂਢ

ਵੀਰਵਾਰ , **01** ਨਵੰਬਰ , **2007** ਨੂੰ **08 : 02 GMT** ਤੱਕ ਦੀਆਂ ਖ਼ਬਰਾਂ

ਕਰਨਾਟਕ ਨੂੰ ਲੈਕੇ ਰਾਜਨੀਤਕ ਜੋਸ਼ ਵਧਿਆ

ਕੁਮਾਰਸਵਾਮੀ ਨੇ ਰਾਜਨੀਤਕ ਚਾਲ ਬਦਲਦੇ ਹੋਏ ਯੇਦਿਉਰੱਪਾ ਨੂੰ ਮੁੱਖਮੰਤਰੀ ਦੇ ਰੁਪ ਵਿੱਚ ਸਮਰਥਨ ਦੇਣ ਦਾ ਫੈਸਲਾ ਕੀਤਾ ਹੈ

ਕਰਨਾਟਕ ਨੂੰ ਲੈਕੇ ਰਾਜਨੀਤਕ ਗਹਿਮਾਗਹਿਮੀ ਤੇਜ਼ ਹੋ ਗਈ ਹੈ ਅਤੇ ਹੁਣ ਇਹ ਮਾਮਲਾ ਦਿੱਲੀ ਆ ਗਿਆ ਹੈ .

ਸੰਭਾਵਨਾ ਹੈ ਕਿ ਜੱਲਦੀ ਹੀ ਕਰਨਾਟਕ ਦੇ ਰਾਜਨੀਤਕ ਭਵਿੱਖ ਦਾ ਕੋਈ ਫੈਸਲਾ ਹੋ ਜਾਵੇਗਾ .

ਬੁੱਧਵਾਰ ਨੂੰ ਜਿੱਥੇ ਭਾਜਪਾ ਦੇ ਉੱਚ ਪੱਧਰ ਪ੍ਰਤਿਨਿੱਧੀ ਮੰਡਲ ਨੇ ਪ੍ਰਧਾਨਮੰਤਰੀ ਮਨਮੋਹਿਨ ਸਿੰਘ ਨਾਲ ਮੁਲਾਕ਼ਾਤ ਕੀਤੀ ਹੈ ਪ੍ਰਧਾਨਮੰਤਰੀ ਅਤੇ ਸੋਨਿਆ ਗਾਂਧੀ ਨੇ ਕਾਂਗਰਸ ਕੋਰਗਰੁਪ ਦੀ ਬੈਠਕ ਵਿੱਚ ਕਰਨਾਟਕ ਦੇ ਰਾਜਨੀਤਕ ਹਾਲਾਤ ਤੇ ਚਰਚਾ ਕੀਤੀ ਹੈ .

ਓਧਰ ਭਾਜਪਾ ਨੇ ਧਮਕੀ ਦਿੱਤੀ ਹੈ ਕਿ ਜੇਕਰ ਭਾਜਪਾ - ਜਨਤਾਦਲ (ਏਸ) ਨੂੰ ਸਰਕਾਰ ਬਣਾਉਣ ਲਈ ਸੱਦਿਆ ਨਹੀਂ ਕੀਤਾ ਗਿਆ ਤਾਂ ਦੋਨਾਂ ਦਲਾਂ ਦੇ **129** ਵਿਧਾਇਕ ਰਾਸ਼ਟਰਪਤੀ ਦੇ ਸਾਹਮਣੇ ਪਰੇਡ ਕਰਣਗੇ .

ਲਿਖਣ ਯੋਗ ਹੈ ਕਿ ਜਨਤਾ ਦਲ (ਏਸ) ਅਤੇ ਭਾਜਪਾ ਦਾ ਗੰਢ-ਜੋੜਾ ਸੱਤ ਅਕਤੂਬਰ ਤੱਕ ਸੱਤਾ ਵਿੱਚ ਸੀ . ਪਰ ਸਮੱਝੌਤੇ ਦੇ ਅਨੁਸਾਰ ਮੁੱਖਮੰਤਰੀ ਬਦਲੇ ਜਾਣ ਦੇ ਝੱਗੜਾ ਦੇ ਚੱਲ ਦੇ ਗੰਢ-ਜੋੜਾ ਟੁੱਟ ਗਿਆ ਅਤੇ ਭਾਜਪਾ ਨੇ ਸਰਕਾਰ ਨਾਲ ਸਮਰਥਨ ਵਾਪਸ ਲੈ ਲਿਆ .

The accuracy of the translation comes out be approx. 95%.

VIII. Conclusion

The present system is translating any complex sentence. The System accuracy is measured up to 95%. This web tool has number of applications in real world. This web tool can be used by Newspaper agencies, any website owner, using email facilty by community of different countries or regions i.e. Writing the email in Hindi and recipient will receive the email in Punjabi. Thus removing the language bars, communication becomes easy in one’s owm language.

REFERENCES

1. Kemal Altintas, Dept. of Computer Engineering, Bilkent University, Ankara, Turkey, “A Machine Translation System Between a Pair of Closely Related Languages”. Internet: http://www.cs.bilkent.edu.tr/~ilyas/PDF/iscis2002.pdf
2. Bharati A., Chaitanya V and Sangal R, "Natural Language processing: A Paninian Perspective", Prentice Hall of India, New Delhi, 1995.
3. Joseph Seasly, “Machine Translation: A Survey of Approaches”, University of Michigan, Ann Arbor, 2003.
4. Durgesh Rao, “Machine Translation in India: A brief survey”, National Centre for Software Technology, Mumbai, 2001. Internet: http://www.eldra.fr/en/rproj/scalla/SCALLA2001Rao.pdf
5. R.M.K. Sinha, R. Jain and A. Jain "Translation from English to Indian Languages: ANGLABHARTI Approach", Proceeding of STRANS-2002, pp. 69-85, 2002.
6. Christopher D. Manning and Hinrich Schutze, “Foundations of Statistical Natural Language Processing”, MIT Press,1999.
7. Manish Sinha, Mahesh Kumar Reddy, Pushpak Bhattacharya, “Hindi Word Sense Disambiguation”. Internet: http://www.cse.iiitb.ac.in/Pb/papers/HindiWSD.pdf
8. R. Canals-Marote, A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, and M.L. Forcada, “The Spanish_ Catalan machine translation system interNOSTRUM”, Internet: http://internostrum.com/docum/iN-MTS.pdf
9. Kemal Altintas, Ilyas Cicekli,”A Machine Translation System Between a Pair of Closely Related Languages”, Internet: http://www.cs.bilkent.edu.tr/~ilyas/PDF/iscis2002.pdf
10. Kevin P. Scannell ,”Machine translation for closely related language pairs” Internet: http://borel.slu.edu/pub/ga2gd.pdf
11. Jan AJIC, Jan HRIC, Vladislav KUBON, “CESILKO – an MT system for closely related languages”, Internet: http://www.cs.ust.hk/acl2000/Demo/03_kubon.pdf