MOVING MACHINE TRANSLATION SYSTEM TO WEB

GURPREET SINGH JOSAN

Dept of IT, RBIEBT, Mohali.
Punjab ZIP140104,India
josangurpreet@rediffmail.com

Abstract

The paper presents an overview of an online system based on Punjabi to Hindi Machine translation system. The implementation of the system is roughly divided into two parts: the client side and the server side. On the client side, issues concerned are development of user interface. As the system accepts input in Unicode, the interface enabling user to enter text in Unicode is needed. On the server side, server handles the requests for the translation of the text written in Punjabi and returns the result to the client. The database serves as a source of language-specific content such as language rules, word and phrase dictionaries. Some web based tools like website translation on the fly, CLIR are developed using machine translation system.

Keywords: Machine Translation, WWW, CLIR

1. Introduction

Computer and Internet are supposed to be a great tool of cultural invasion and play a vital role in spreading the influence of western and particularly English culture in developing countries. None other than the technology itself can counter this effect. Enabling users to process and communicate data in their regional languages proves to be great tool to counter the effect of technology on civilizations. The research and hence the tools developed as a result of research must reach en masse for the effective utilization of technology. It can help in blurring the barriers among different nations and society and boost the process of world globalization. In a multicultural and multilingual country like India, where unity in diversity prevails, technology plays a vital role to further enhance the bond among people. Language barrier is a key issue for a wide interaction among various cultures in India. Natural language processing and its sub fields got attention of researchers in recent times. Machine translation is now mature enough to provide some useful results. Such system can be more useful if they can be reached by users anytime. This kind of availability can be provided by internet very efficiently and effectively. This paper discusses the issues in moving a standalone Punjabi to Hindi machine translation system on web.

The power of the machine translation system will not be envisaged until we put it on the Web for every one's use. Among the various aims of this research, one is to reach the masses at grassroots level with our system and break the language barrier enabling them to communicate with one another and retrieve the vast amount of information available in the target language on the internet. It is this idea of convenience of communicating over the Web in one's own mother tongue that has inspired the formation of tools which exploit the machine translation modules for the realistic use by the masses.

2. Previous Work

The free online translation service has been available in the market since 1994 (Somer, Gaspari, & Ana, 2006). From its availability, it has changed the view of general public about the machine translation and played a vital role in shaping and developing the technology. Due to the features of World Wide Web like easily accessibility, few companies have attempted to make machine translation available on web (AltaVista, 1999; FreeTranslation, 1999; InterTran, 1999). Earlier systems cannot be used for real-time, speech-to-speech communication with translation and primarily used for amusing oneself. Most of the early attempts to make system online aims at developing interactive web based system for ts improvement. Hogan C. and Frederking R., 2009, implemented an interactive, Web-based, chat-style machine translation system, Supporting speech recognition and synthesis, local- or third party correction of speech recognition and machine translation output, and online learning. The underlying client-server architecture, implemented in Java, provides remote, distributed computation for the translation and speech subsystems. Hettige and Karunananda, 2008, reports on the development of on-demand machine translation of selected texts from an English document to Sinhala. Their work is a web-based extension of English to Sinhala MT system. Romero et al., 2009, discuss a web-based natural language processing systems, where the user's feedback has shown to improve system accuracy, and increase both system ergonomics and user's acceptability. Ortiz-Mart´ınez et al., 2010, proposed interactive approach as an alternative to post-editing the output of a machine translation system. In this approach, the user's feedback is used to validate or to correct parts of the system output that allow the generation of improved versions of the

rest of the output. Way and Gough, 2003, had developed an example-based machine translation (EBMT) system that uses the World Wide Web for two different purposes: First, we populate the system's memory with translations gathered from rule-based MT systems located on the Web. They presented an EBMT system based on the marker hypothesis that uses post hoc validation and correction via the Web. Despite the fact that the output from on-line MT systems is often faulty, they demonstrate in a number of experiments that when used to seed the memories of an EBMT system, they can in fact prove useful in generating translations of high quality in a robust fashion.

3. Aim of tool Development

A machine translation system serves as an instrument for reading and communicating such that users can obtain a general idea of text written in the language other than their own. The aim of this work is to develop a Machine Translation Web Service that can be accessed online to help users to translate their text from Punjabi to Hindi. A web based utility has also been developed that will convert the web site in Punjabi to Hindi for the readers who know Hindi only and can't read Punjabi. A cross language information retrieval system also has been developed where a user can put his query in Punjabi language and system will translate the query in Hindi and then extract the information from Hindi webpages. Although online translation services are being provided by many companies for quite a long time, none of those involves the Punjabi-Hindi language pair. Attention has been paid towards the development of better quality online translation service. Next sections discuss all these tools in detail.

4. Translation System

A Punjabi to Hindi Translation System has been developed (Josan & Lehal, 2008). The system is a direct machine translation system based upon exploitation of syntactic similarities between more or less related natural languages like Punjabi and Hindi. In this system, words from source language are chosen; their equivalents in target language are found out from the lexicon and are replaced to get target language. The source text is passed through various pre processing phase and output is also passed through a post processing phase. The system uses various resources like root word lexicon, inflectional form lexicon, ambiguous word lexicon and bi and tri gram tables etc. The text passed through various steps like text normalization, tokenization, Translation, post processing etc. The translation module also takes care of various syntactic structures in a sentence like repetitive construct, rhyming reduplication, Named entity recognition, transliteration and ambiguity resolution. The output of translation module is passed through post processing phase where rules are applied to remove some discrepancies and make output more grammatical.

5. Moving System on Web

The implementation is roughly divided in two parts: the client side and the server side. The client side contains the translation engine which accepts request from the client in Unicode, translate it in target language using various lexical resources and then send the response back to the client. The lexical resources are placed on the server. Client side is responsible for presenting an interface to end user enabling him to communicate with server. The interface includes options to write text in Unicode and to display the results of translation engine. The system described in section 4 above is placed at server. Following sections discuss different tools developed on client side.

5.1. *Text Translation Web Page*

This is the webpage where the interaction between the service and web user takes place. It allows the user to pass the Punjabi text to the web server for translation. The input text can be entered in a textbox provided there. The foremost issue in a web based translation system is to enter text in regional language. The translation system is Unicode based and accept the input in Unicode format. We need to design a web based system that is capable of taking input in Unicode or we need to develop a module that accept input in font encoded form and convert it to Unicode before sending it to translation engine. Three options are provided to the end user. The data can be entered in a given textbox through standard keyboard. For typing from keyboard, a module is developed in JavaScript which enables a user to enter the Punjabi text in Unicode format by masking the keys of keyboard. This module uses the key map of AnmolLipi Font which is phonetic based. Key maps of other popular fonts can also be added easily. Intuitive text suggestion is also provided to the user as he/she types in text box. Lists of words appear above text box that suggests the words starting from the alphabet that a user starts to type. User can select the desired word from this list. It also alleviates the problem spelling errors and thus helps in improving the output. Second option is on screen keyboard. The user is provided with onscreen

keyboard and can select the desired letter by pressing the buttons of keyboard with mouse. For large amount of data, users also have the option to upload the file in Unicode format. The file then read by the system and the text in file will appear in the given text box.

When a user submits the text for translation, it is passed on to the server. At server side, an object of translation class is instantiated and the whole text is passed to the translate function of this object as argument. This function returns the translated text which in turn is passed on to the client and appears in the output box on client side. Following is the screen shot of this page.

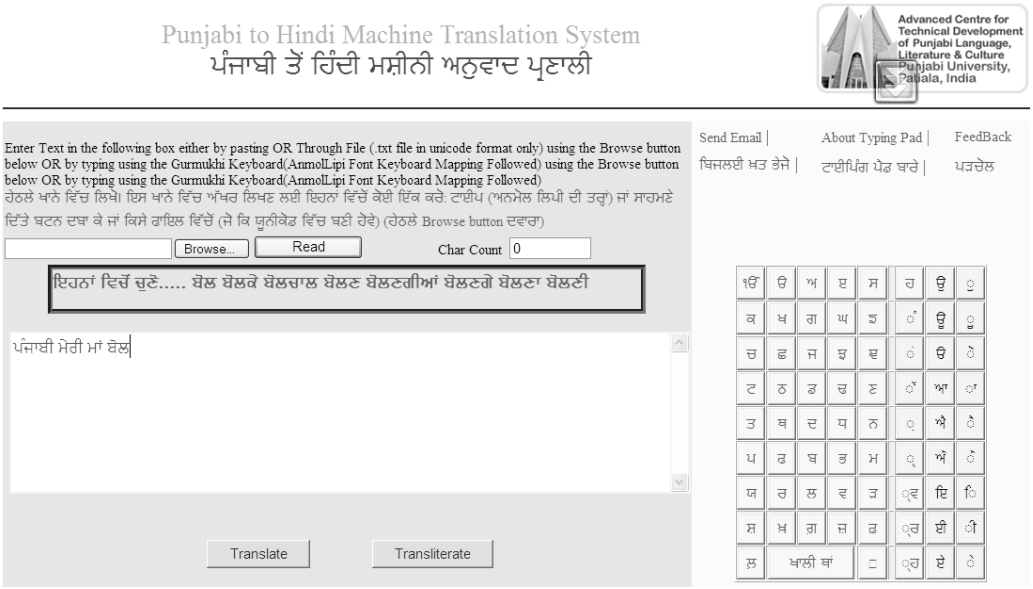

Fig 1 User interface for text translation

5.2. *Website Translation service*

Apart from translating simple text, a more useful application is to translate a whole web page on the fly. For this, a webpage is created where a web user can submit a URL of Punjabi websites of his interest for translation. It consists of following modules:

Retrieving and Parsing HTML Page
Conversion of Non Unicode text to Unicode format
Translating
Combining the translation Unit with HTML Codes
Altering the links in webpage
Displaying the result

The whole architecture is shown in fig. 2

Retrieving and Parsing HTML Page

The URL of Punjabi webpage passed by the web user is used to generate an HTTP Request. This request is sent to the server hosting webpage. Against this request, the remote server responds and this response is fetched and saved in a Buffer at translation server. Then a HTML parser will commence the process of analyzing and extracting plain text within formatting tag in that webpage. During analysis, it notes down the most recent font used from the attributes of font tag.

Conversion of Non Unicode text to Unicode format

If the text in website is not in Unicode format, then it must be converted to Unicode format before passing to translation service. Detecting the font used in a particular HTML tag is a non trivial task. There are many ways to specify fonts: cascading style sheets (CSS), style attributes and font tags. Any imbedded tags inherit the styling of parent tags, unless they specifically override the style. The font in our system is detected by checking the most recent font tag. If no font name is described, then it is assumed to be Unicode. Once font name is detected, the whole extracted text is passed to the conversion unit where it gets converted into Unicode format.

Translating

The converted text is then passed to the translating unit for translation. The text is passed through various modules of translation unit and the target text is produced.

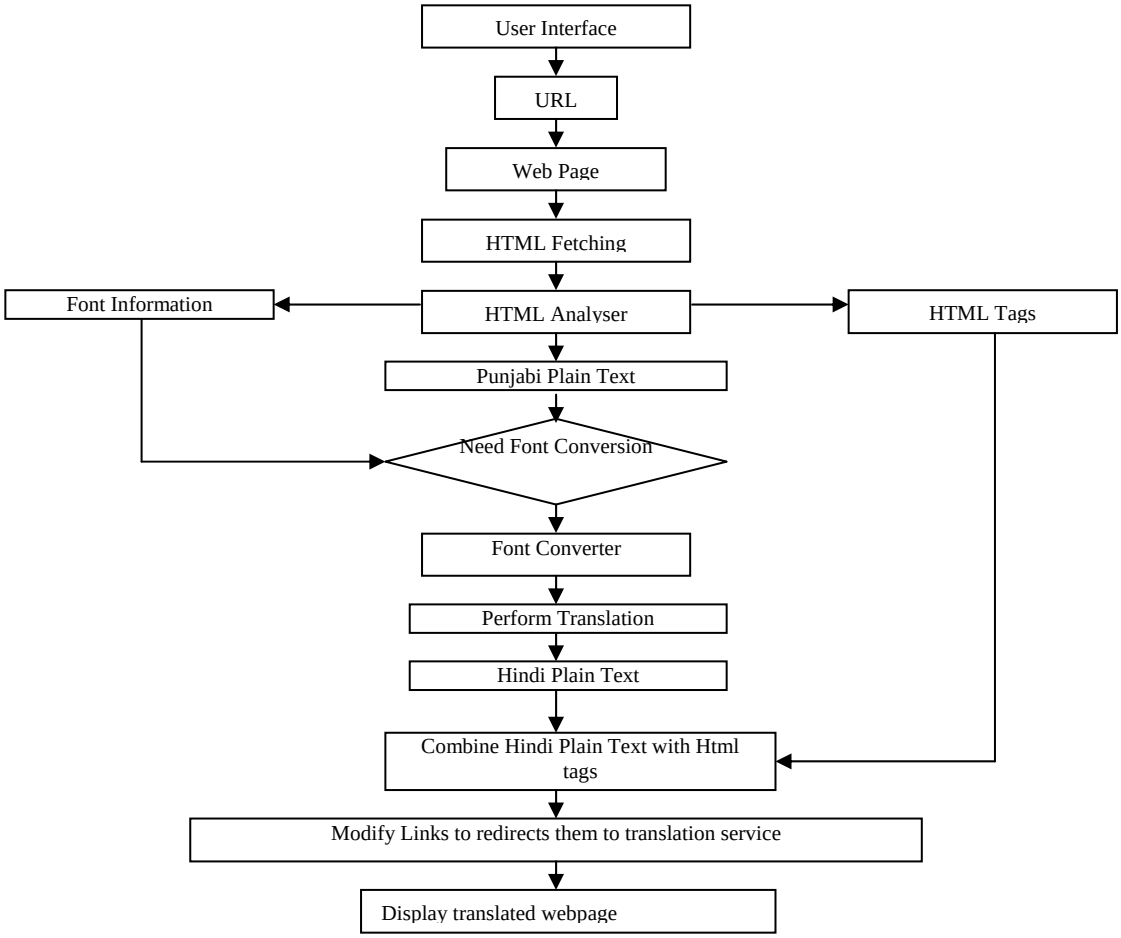

Fig 2 Architecture of Web Translation Service

Combining the translation Unit with HTML Codes

The text produced by Translating unit is again fixed at its appropriate position in the HTML document. While replacing the source text with target text, the font name attribute of each text is changed to Anmol Unicode MS.

Altering the links in webpage

At the final stage all the links in webpage are replaced, so that the next links must redirect the request through our translation service. By this step, user does not need to enter URLs or take any other action if he/she wants to translate the linked page. He/She simply needs to click on the given link.

Displaying the result

The translated webpage is then forwarded to the client in the same format in which the original page had appeared.

Screen Shots are as follows:

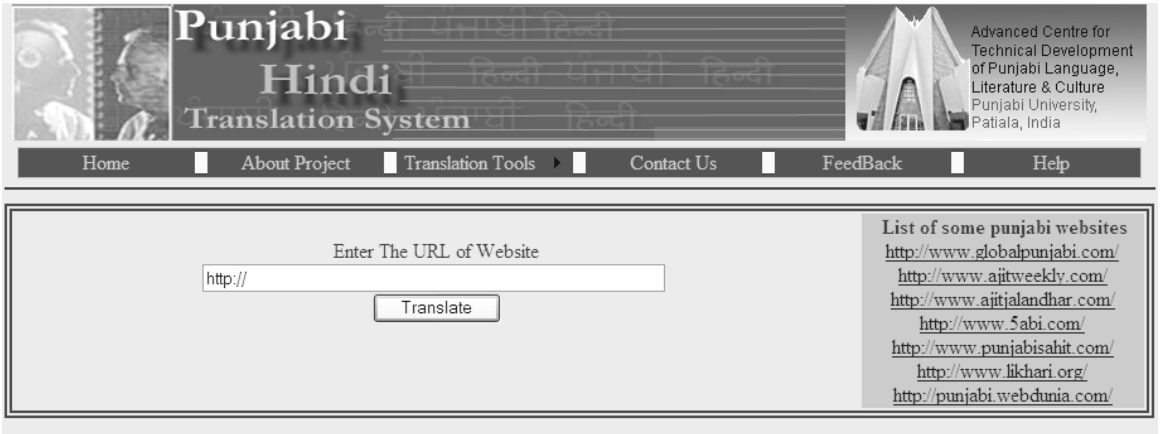

Fig 3 User interface for website translation

Fig 4 Original Web site http://www.ajitweekly.com/ on 31 March 2011

Fig 5 Translated Web Site http:// www.ajitweekly.com/

5.3. *Cross Language Information Retrieval*

Before the decade of 1990, the machine translation systems were widely used as standalone systems which translate text and sometimes couple at most with text processing systems. With the advancements in technology, a vast amount of information in different languages is available on the internet. As put by Peterand Picchi, 1997 the authors of these resources want their documents to be translated and retrieved with greater accuracy, and thus to have a better chance to reach the right kind of readers. On the other hand, the readers wish to find the best matched resources without wasting time in non-related documents. Moreover, an appearance of these resources in more than one language is becoming common particularly:

— in countries with more than one natural language

— in countries where both the national language and English are commonly used for scientific and technical documentation

— in multinational companies

This information can be retrieved and utilized by the end users by integrating the MT system with other text processing services such as text summarization, information retrieval, and web access. It enables the web user to perform cross language information retrieval from the internet.

Cross-Language Information Retrieval (CLIR) enables users to construct queries in one language and search the documents in another language. The Cross-Language Text and Speech Retrieval challenge termed as the "Grand Challenge" statement and discussed in the wrap-up panel session of 1997 AAAI Spring Symposium is as follows:

> *Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified.*

Thus CLIR is naturally associated with MT (Machine Translation) and IR (Information Retrieval). One way to tackle the multilingual information retrieval problem is to translate all the target language text into source language text and then perform monolingual search on the translated text. Without better machines and high speed/quality MT, we can rule out the practical application of this approach for the web. We, therefore, adopt the query translation approach. To translate user queries from source languages to target languages, we need multilingual/bilingual transfer dictionaries or corpora (parallel or non-parallel).

CLIR Web Interface for Punjabi Language

Hindi being the national language of India is widely used in numbers of web pages. The government website contains lot of information regarding government policies and rules & regulations. Besides this, other websites like newspapers site, tour and excursion planning sites etc also has useful information for a person living in Punjab and using Punjabi Language. To access all this information a web based prototype model has been developed in which a user can enter the text in Punjabi. This text is translated to Hindi by using the translation service. The translated text is then posted for query on Google search engine. The search engine retrieves the related websites and the results are represented to the end user. The whole architecture is as shown in fig 6 and a snapshot is given in fig 7.

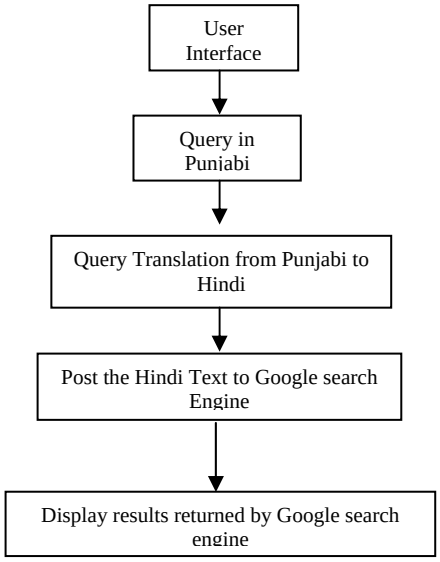

Fig 6 CLIR Architecture

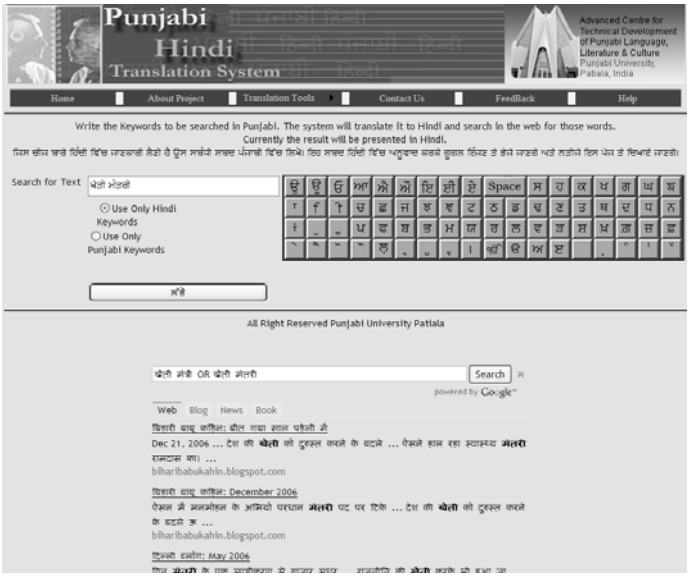

Fig 7 Interface of cross language information retrieval

6. Conclusion

In this paper, the issues pertaining to moving a machine translation system to web are discussed. Major concerns during the development are the convenience to the end user for text input in regional language. An interface is developed which is user friendly and suggests user the correct spelling as he types. Online system can expand the opportunity to access the MT system by general public, and also to provide a good feedback for evaluation and further improvement to the system. It can also provide a foundation for machine translation from Punjabi to any language. Our further research will be towards the following fields. The system will be moved on hand held devices and mobile phones for the wider reach and usage.

References

[1] AltaVista. 1999. Babel Fish: A SYSTRAN translation system, http://babelfish.altavista.com/.
[2] FreeTranslation. 1999. FreeTranslation: A Transparent language translation system, http://www.freetranslation.com/.
[3] Hettige B. , Karunananda A. S., 2008, “Web-based English to Sinhala Selected Texts Translation system”, In proceedings of Sri Lanka Association for Artificial Intelligence, Fifth Annual Sessions 2008
[4] Hogan C. and Frederking R., “WebDIPLOMAT: A Web-Based Interactive Machine Translation System” Proceedings of the 18th conference on Computational linguistics - Volume 2. PP 1041 – 1045, 2000.
[5] InterTran. 1999. An InterTran translation system. http://www.airsho.com/transLator3.htm
[6] Josan G., Lehal G., 2008, “A Punjabi To Hindi Machine Translation System”, In proceedings of Computational Linguistic (COLING- 2008), Manchester.
[7] Ortiz-Mart´ınez D., Leiva L. A., Alabau V., Casacuberta F., 2010, “Interactive Machine Translation using a Web-based Architecture”, In proceedings of IUI 2010.
[8] Peter, C. and Picchi, E. (1997). "Across Language, Across Cultures". D-Lib Magazine, May. URL:http://www.dlib.org/dlib/may97/peters/05peters.html
[9] Romero V., Leiva L. A., Toselli A. H., and Vidal E., 2009. “Interactive multimodal transcription of text images using a web-based demo system”, In Proceedings of IUI, 2009.
[10] *Way A., Gough N., 2003, “ wEBMT*: Developing and Validating an Example-Based Machine Translation System Using the World Wide Web”, Computational Linguistic, Volume 29, Number 3, Association for Computational Linguistics, 2003, 421-457.