Sampark: Machine Translation Systems for Indian Languages

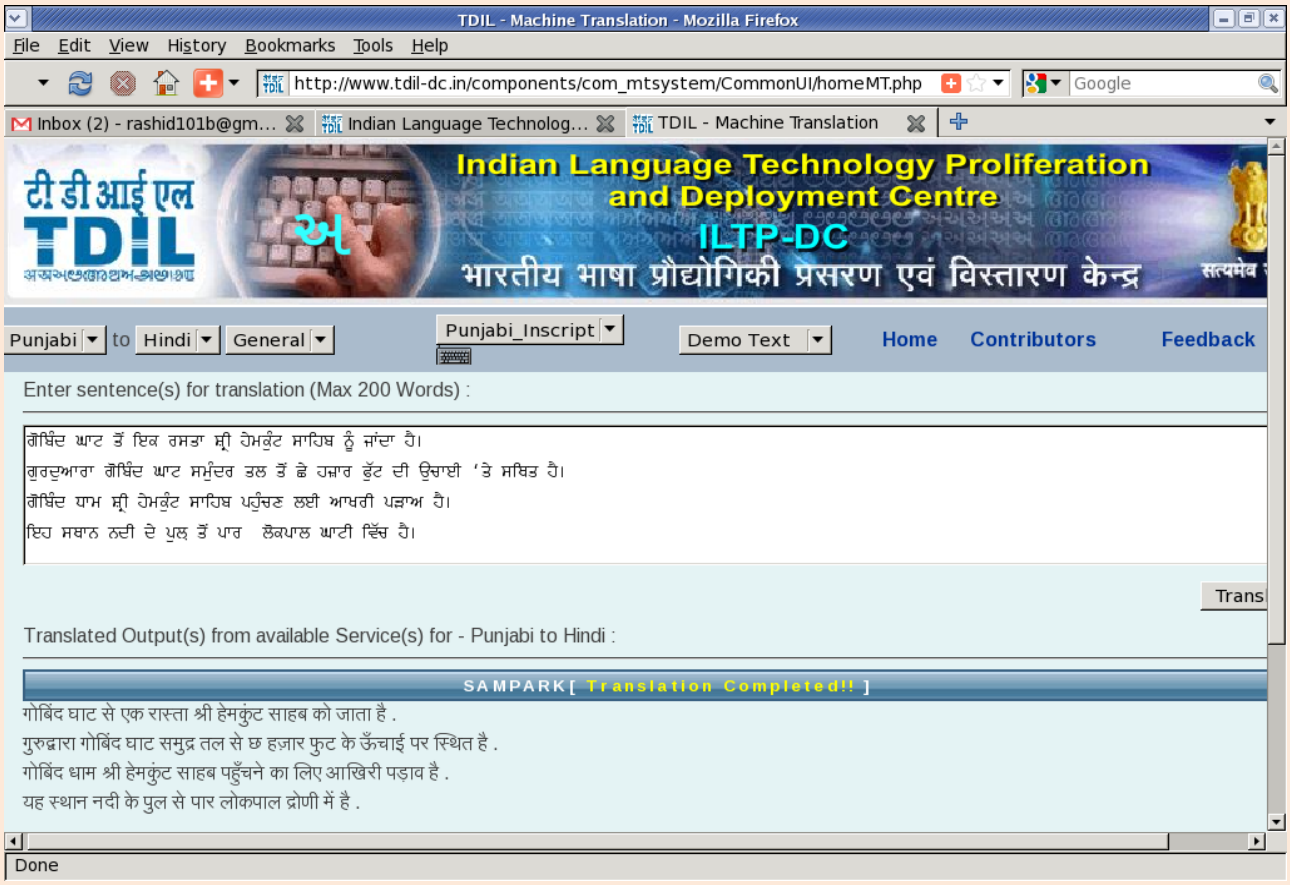
Sampark is an Indian languages translation system that aims at instant, high quality translation between 18 Indian language pairs. Sampark MT has been developed by a consortium of 11 Indian academic and research institutes, led by IIIT Hyderabad with the support of "Technology Development for Indian Languages" (TDIL) Programme of the Department of Electronics & Information Technology (DeitY), Government of India.

The Indian Languages to Indian Languages Mt system is being developed for the 18 language pairs i.e. Punjabi – Hindi –Punjabi, Telugu – Tamil – Telugu, Urdu – Hindi – Urdu, Hindi – Telugu – Hindi, Marathi – Hindi – Marathi, Bengali – Hindi – Bengali, Tamil – Hindi – Tamil, Kannada – Hindi – Kannada, Malayalam – Tamil – Malayalam.

Sampark systems for the following language pairs - Punjabi to Hindi, Hindi to Punjabi, Urdu to Hindi, and Telugu to Tamil - have been released on TDIL Data centre www.tdil-dc.in for public use and feedback.

IL-IL MT System is based on the Computational Paninian Grammar (CPG), which works very well for free word order languages, and, particularly, for Indian languages. It is a hybrid system using Paninian rule based approach and statistical machine learning on tagged texts.

Built on blackboard architecture to provide a common framework for integrating different modules. The architecture helps minimize complexity, besides achieve modularity, transparency, scalability, and flexibility in development.



A common GUI for the ease of users is provided. The screens shows a glimpse of Punjabi-hindi MT System

Sampark - Technical Features

Building an MT system requires analyzing the source language text to understand the meaning, performing a dictionary lookup and structure transfer, and finally generating the target language text. There are two key parameters to judge the quality of MT output - comprehensibility (with respect to the meaning in the source text) and fluency (naturalness in the target language text). ILMT is currently focusing on comprehensibility, and so the output text may not be very natural, at places. The complexity and diversity of languages pose a host of immense computational challenges in building automatic translation or machine translation (MT) systems.

- Accepts sentence input on-line or from file.
- Sampark translates the input text (in Unicode), one sentence at a time.
- The Sampark MT system has been built for general purpose text and is not limited to any particular subject domain.
- Sampark can be customized for use in other domains, with a little effort.
- Sampark offers better quality output if the input text conforms to standard language rather than being in an informal style.

Possible Use

The sampark systems have reached a stage of development where some of these can be made available to the users. There are various deployment possibilities.

A) The systems could be made available on the web as a service for general purpose use. Here the idea is that the systems are made available to general users. They can copy–paste a text, they wish to read, and get instant translation on demand.

B) MT as a service may be provided on some popular websites such as regional language newspapers' web pages, or on a search engine's web page or on web pages in a domain such as tourism, etc. People who access the page would have the MT system as a service and could get the web page's content translated into a language of their choice.

C) The systems can be used by making it available as a translator's help desk to the professional translators both individuals as well organizations.

Consortia Leader:
Dr. Rajeev Sangal, IIIT Hyderabad
Consortia Members:
IIT Bombay, IIT Kharagpur, C-DAC Noida, University of Hyderabad, Jadavpur University, Anna University-KBC Research Centre, Tamil University, IIIT Allahabad, IISc Bangalore, IIITM-Thiruvanantapuram
Support:
Technology Development for Indian Languages Programme
Department of Electronics & Information Technology
Government of India