

Language Translator

Ekta Ahuja, Karan Kashyap

Abstract— The aim of this project is to find out the problem areas in the translation done by Bing, suggest a method to improve the language translation and build a translator that supports accurate English to Hindi translation. The English input can be of textual, voice or image form and the Hindi output will be in the textual format.. With our research we aim to suggest a way to improve the already available popular translation engines. For this purpose we have used the method of hybrid machine translation. Hybrid machine translation is a method of machine translation that is characterized by the use of multiple machine translation approaches within a single machine translation system. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy.

Keywords—translator, Bing, textual, hybrid.

I. INTRODUCTION

Nowadays, language translation and interpreting services are useful in all spheres of daily life, like to help a student to enroll into a new school, or to interpret for a corporate executive who has just purchased a new business in India. Translation can play a number of different roles such as a ‘unifying’ or constructing new words but also, most crucially, as a source of knowledge about foreign, lesser known cultures. Thus, translation is not merely a linguistic process, but can also make a political and social impact.

II. PROPOSED APPROACH

A. Rule-Based Machine Translation

Rule-Based Machine Translation systems use large collections of rules that are manually developed over time by human experts. They basically map structures from the source to the target language. Rule-based MT requires manually building extensive translation lexicons. Users may often edit or manipulate them for better translations. The human skills invested in the development of rule-based systems have the potential of delivering fairly good automated translations with predictable results. It is possible to adjust the errors safely through changes to user dictionaries. However, as it involves significant manual labor, rule-based systems can be quite costly and time consuming to implement as well as to maintain. But as rules are added and updated, these systems have the potential of generating ambiguity which can cause significant degradation of translation output quality over a period of time.

B. Statistical Machine Translation

Statistical Machine Translation systems make use of computer algorithms that find out many possible ways of connecting smaller pieces of text together, in order to produce a best translation. Statistical Machine translation basically translates words and phrases along with their statistical likelihood. These are learned automatically from previously translated text, creating a bilingual “database” of translations. A program referred as decoder matches the source code and phrases with the translation databases and searches for all possible translation combinations. An algorithm is then used to select the best translation out of the millions of possible translations and output it. The main advantage of statistical systems is the high levels of automation involved in developing as well as maintaining these systems. As machine-learning algorithms are used, deploying new statistical MT systems is greatly faster and more cost efficient as compared to older technologies. Also, it is easy to update this system that leads to improved performance and increased translation coverage over time. However, these translation systems have their shortcomings as well. As automated machine learning techniques are used it becomes difficult to control the learning process and identify the source of potential errors. These systems are limited in addressing language structure when translating from source to target language.

C. Hybrid Machine Translation

Hybrid machine translation is a method of machine translation that is characterized by the use of multiple machine translation approaches within a single machine translation system. The motivation for developing hybrid machine translation systems stems from the failure of any single technique to achieve a satisfactory level of accuracy. In this hybrid approach, the translation is performed using a Rule-Based MT and a Statistical MT is then used in order to adjust the output. In what seems to be a theoretically sensible approach, hybrid solutions intend to combine the advantages of the individual approaches to achieve an overall better translation. The approach is most useful to address one of Rule-Based MT greatest challenges – translation ambiguity. When a word/phrase can have more than one meaning, statistics can help identify the most suitable option. Hybrid MT systems provide quality improvement benefits over each individual approach.

III. REQUIREMENTS

A. Functional Requirements

The system must translate the given input text into the language specified by the user. Our module is based on the language Hindi. The translator should provide an additional feature wherein the translated text can be shared on the various social and professional platforms. It must allow three important features.

Manuscript published on 30 October 2014.

* Correspondence Author (s)

Ekta Ahuja, Computer Science, Vivekanand Education Society’s Institute of Technology, Mumbai, India

Karan Kashyap, Computer Science, Vivekanand Education Society’s Institute of Technology, Mumbai, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](http://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Text to Text.
Speech to Text.
Image to Text.

B. Non Functional Requirements

Loading of a user interface element in a translated format should take no longer than the acceptable standard. Currently this is set at a second to load the page. For the smooth and efficient working of the system, internet connection is must. In order to enjoy the benefits of sharing feature, device must contain the Hindi keyboard application.

IV. IMPLEMENTATION

Our application is build using some available softwares and APIs. For translation, we have made use of Bing API provided for the application development by Microsoft Corporation. And for text recognition on the clicked image, we've made use of the tesseract software. Out of all the softwares available, tesseract is the most accurate one. Because it stores all the possible patterns that a specific alphabet can take. So, it gives the best results as far as text recognition from a image is concerned. Apart from these softwares, we've build our application with the help of android developers tools like The Android Developer Tools (ADT) plug-in for Eclipse provides a professional-grade development environment for building Android apps. It's a full Java IDE with advanced features to help you build, test, debug, and package your Android apps Our Module in the app makes use of Hybrid Machine Translation for the accurate implementation of Eng-Hindi Translation when pronouns are used for gender recognition.

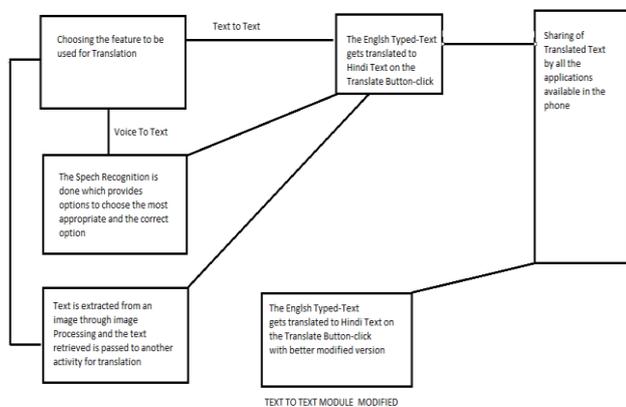


Figure 1: System Design.

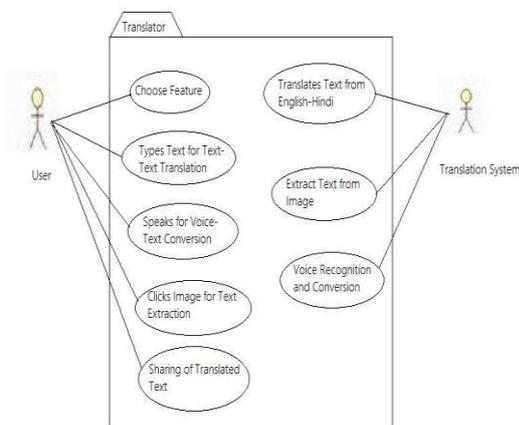


Figure 1: System Design.

V. OBSERVATION AND ANALYSIS

A. Bing Translator

Input is: He sings song.

Statistical analysis of 'sings' is done as follows

| | |
|------------------------|----------------|
| sita <u>sings</u> song | सीतगीतगातीहै |
| ram <u>sings</u> song | रामगीतगातीहै |
| riya <u>sings</u> song | रियागीतगातीहै |
| he <u>sings</u> song | वहगानागातीहै |
| she <u>sings</u> song | वहगानागातीहै |
| anjali <u>sings</u> | अंजलिगीतगातीहै |
| song | वहगानागातीहै |
| she <u>sings</u> song | वहगानागातीहै |

According to statistics sings->गाती

If input statement is

He sings song.

Output is: वहगानागातीहै!

A. Our Module

Input is: He sings song.

In our module if statement contains pronoun then statistical analysis is carried out for both pronoun and verb

According to statistics sings->गात

If input statement is

| | |
|----------------------|----------------|
| sita sings song | सीतगीतगातीहै |
| ram sings song | रामगीतगाताहै |
| riya sings song | रियागीतगातीहै |
| <u>he</u> sings song | वहगानागातीहै |
| she sings song | वहगानागातीहै |
| anjali sings song | अंजलिगीतगातीहै |
| she sings song | वहगानागातीहै |

He sings song.

Output is:

वहगानागाताहै!

VI. SCOPE AND LIMITATION

Our module has been developed for the sentences of the form 'Subject Object Verb'. It needs to be expanded to other sentence formats to improve efficiency and increase the usage scope. The significance and relevance of translation in our daily life is multidimensional and extensive. It is through translation we know about all the developments in communication and technology and keep abreast of the latest discoveries in the various fields of knowledge, and also have access through translation to the literature of several languages and to the different events happening in the world. The future scope of our project at the major scale comes into significance when translation between English to Hindi is done using sentences of the type Subject-Object-Verb (in English) where subject is name of the person or any pronoun that addresses a gender (he/she).



ACKNOWLEDGEMENT

We are grateful to Vivekanand Education Society's Institute of Technology for providing continuous support throughout the work.

REFERENCES

- [1] RAY SMITH, AN OVERVIEW OF THE TESSERACT OCR ENGINE, PROC. NINTH INT. CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR), IEEE COMPUTER SOCIETY (2007), PP. 629-633
- [2] KEVIN KNIGHT, A STATISTICAL MT TUTORIAL WORKBOOK, JHU SUMMER WORKSHOP, APRIL 30, 1999
- [3] A. Srivastav, English to Hindi Machine Translator Using GMT and RBMT Approach Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India

Ekta Ahuja is a Computer Engineering student from VESIT, Mumbai. She has made various projects using Python and believes that technology and management should go hand in hand

Karan Kashyap is a Computer Engineering student from VESIT, Mumbai. He has made projects in ASP .NET and Java. He has keen interest in technology in general and looks at himself dealing with finance and analysis in related fields