Smart Chat with Translator (English ↔ Hindi) Domain Natural Language Processing

Tehsin Khan, Samiksha Patwari, Sadhna Bathe and Prof. Nasim Shah

Abstract— Smart Chat (SC) is a form of real-time Communication between two or more people based on typed text. The text is conveyed via computers connected over a network such as the Internet. Today chat system is an important technology for most Internet users all over the world. This article describes a new development in the branch of chat systems that allows to translate typed and received messages from one language to another. In this paper the solution will help the people who speaks in different languages to communicate with each other through on-line messengers located in different countries.

Keywords--- Attribute Logic Engine (ALE), Morphological Segmentation (MS), Named Entity Recognition (NER), Parsing, Smart Chat (SC), Word Sense Disambiguation (WSD)

I. INTRODUCTION

SMART Chat(SC) is a form of real-time communication between two or more people based on typed text. The text is conveyed via computers connected over a network such as the Internet. SC offers real-time communication and allows easy collaboration, which might be considered more akin to genuine conversation than email's "letter" format. In contrast to e-mail, the parties know whether the peer is available. SC makes use of the functionality to implement a translating chat room. When a user connects to the SC, they register their username and language code. When a user sends a chat message to the other user the SC takes that message and translates the message from the user's language in to the languages of the other user connected to the SC. Then, it sends out the translated messages to the other user. This effectively means that each person sees all messages in their own language, regardless of the original language the messages were typed in. It also supports user list, words filtering, emoji, sharing of images and data and more.

Some of the NLP tools such as ALE, SYNTACTICA and NL Builder etc are referred to implement NLP tasks like morphological segmentation, NER, parts of speech tagging, parsing and WSD. ALE is an environment that integrates phrase structure parsing and constraint logic programming with typed feature structures. SYNTACTICA is a system for grammar development with a simple graphical user interface.

It is intended for use in introductory syntax classes, or introductory linguistics classes with a syntax component. NL Builder (TM) may be used to develop NLP applications or experiment with various linguistic components. It consists of a tokenizer, a dictionary, a morphological analyzer, a parser, a semantic interpreter, a semantic network KRL, lexical acquisition tools, "C" hooks, and a debugger. None of the chatting applications provide language translation facility. This paper combines the features of both, so that one can write text in their own language without knowing the other user's language.

II. LITERATURE SURVEY

Andrew Rutkas presents Instant Messaging (IM) which is a form of real-time communication between two or more people[1]. This system is an important technology for most Internet users all over the world. Instant messaging systems allows to translate typed and received messages from one language to another. Most systems allow the user to set an online status or away message so peers are notified when the user is available, busy, or from the computer. On the other hand, people are not forced to reply immediately to incoming messages. Rina Damdool, Urmila Shrawankar presents a Statistical Machine Translation(SMT) through which sentences written with short forms, misspelled words and chatting slang can be corrected[6]. Given a source-language (e.g., Short message) sentence, the problem of machine translation is to automatically produce a target-language (e.g., Long form English) translation, to be used by the young generation for messaging. The main goal is to analyze the improvement in efficiency as the size of bilingual corpus increases. Machine learning and translation systems, dictionary and textbook preparations, patent and reference searches, and various information retrieval systems are the main applications of the project. Mehdi Mohammadi presents an EBMT system where the input sentence is matched to some units of text and the translations of these text units are recombined to generate a translation for the input sentence. [3]. For a typical EBMT system, an example set of prepared translation pairs is necessary. It is also needed to have a matching mechanism to retrieve the best example that matches an input sentence. But one of the most difficult steps in the EBMT approach is the recombination in which the translation of matched examples would be mixed together in such a way that the generated output sentence would be in a correct state of grammar and meaning. Kevin Metai, Arturo Trujillo presents A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns. [4] These consist of possibly discontinuous text fragments, with the bilingual relationship.
between the text fragments and the discontinuities between them made explicit. The patterns are extracted from a bilingual parallel corpus aligned at the sentence level, without the need for linguistic analysis, and are used to build a translation memory database which is intended for use in a machine aided human translation (MAHT) setting, such as a translator’s workbench (TWB). The patterns extracted could also form the basis for example-based machine translation (EBMT) without the need for complex linguistic or statistical processing. Given a TM database made up of our concept of translation patterns and a SL input string, relevant translation patterns combine to form TL translations as suggestions to the translator. Victor H. Yngve presents A word-for-word translation consists of merely substituting for each word of one language a word or words from the other language[5]. The word order is preserved. Of course, the machine would deal only with the written form of the languages, the input being from a keyboard and the output from a printer. R.Ravi and S.Kailasam presents a Sentence-for-sentence translation[6]. The idea of NLP is to design and build a computer system that will analyze, understand and generate natural human-languages. It studies the problems of automated generation and understanding of natural human languages. Natural language generation systems convert source human language from computer databases into different target human language. Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

III. SYSTEM DESCRIPTION
This Smart Chat application first will ask for login into account and once login is successfully done it will show the chat window. If any new user wants to register it will provide registration form and will assign register id to that user. Later on he/she will be required to use his/her login id or password to facilitate the chat application. Then he/she will select another user with whom he/she wants to chats. Then both the users will select the language and then they will start communication to each other. Then our application will convert that text into selected language and then text will shown to recipient’s window. This will remove the communication barrier between people. This chat system also provide the facility of file sharing, image sharing etc and also user can set his/her profile.

IV. SYSTEM DIAGRAM
The system design will be as shown in the figure below:

V. ALGORITHM
Steps:
1. Source text dictionary lookup and morphological analysis
2. Identification of homographs
3. Identification of compound nouns
4. Identification of noun and verb phrases
5. Processing of idioms
6. Processing of prepositions
7. Subject-predicate identification
8. Syntactic ambiguity identification
9. Synthesis and morphological processing of target text
10. Rearrangement of words and phrases in target text

Comparison between Proposed System and Google Translator

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>SMART CHAT</th>
<th>GOOGLE TRANSLATOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFFICIENCY</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>CHAT SUPPORT</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>LANGUAGE SUPPORT</td>
<td>LESS</td>
<td>MORE</td>
</tr>
<tr>
<td>SLANG WORD SUPPORT</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>

VI. IMPLEMENTATION PLAN

- **Software Requirement**
  We will use Microsoft Visual Studio2010 for developing our smart chat application because it combines different inbuilt tools and it also supports many design languages and different forms.

- **Hardware Requirement**
  It requires 512kb of RAM and Pentium p4.
Database

We will use Oracle11G to create a database for registry of users whose are logged into the system. The dictionaries for both the languages source language and target language (English and Hindi) are provided to map the words. Linguistic and lexical rules are taken into consideration.

Tools

ALE (Attribute Logic Engine) Conc, ENGLEX, FLEX (Fast Lexical Analyzer Generator), FONOL, Grammar Workbench, KGEN, LINK, Lotec, PC-KIMMO, SAX (Sequential Analyzer for syntax and semantics), SYNTACTICA, NL Builder (TM), VisualText (R).

VII. Future Scope

This system can be further implemented as an android or an iphone app and can also enhance the feature of supporting many more languages.

VIII. Conclusion

This solution brings a new scope in online messaging and internet communication and demonstrates a powerful and fast language translation engine. It can be used as individual internet users and also by companies.

REFERENCES

[1] Language Translation Module for Instant Messaging Systems by Andrew Rutkas
[5] Sentence-for-sentence translation by Victor H. Yngve