Abstract

Machine Translation is a process by which a text in one language is automatically translated to a text of another language. There are two main approaches to build a machine translation system, namely, *Rule Based Machine Translation* and *Statistical Machine Translation*. A machine translation system which is not perfect often returns erroneous sentences. It is possible to improve translation by correcting the errors in the texts generated from such system.

In this thesis, we use different methods to improve the quality of the sentences generated from a baseline Bengali to Hindi machine translation system. This is done by using additional linguistic modules, incorporating some Bengali to Hindi transfer grammar rules and modifying some existing modules. Based on studying errors observed in the translation of a corpus, we have identified the possible sources of the errors in the system. We categorize the sources of errors into word level, chunk level, sentence level and discourse level. By studying these sources of errors, we identify the new modules that need to be integrated, the transfer grammar rules and the modules that need to be modified.

The relations between different words play important roles in machine translation. The important semantic roles have been identified and a *dependency parser* has been built for identifying and classifying these dependencies.

In the baseline system the named entity words are translated when they are found in the dictionary or they are not recognized. We incorporate a *Named Entity Recognizer* (NER) to identify the named entities so that they can be transliterated.

In Bengali clauses the copula verbs in present tense and positive polarity are often dropped. The clauses need to be identified correctly for their correct translations. Sometimes the Bengali clauses are separated by a referent, a co-referent or a conjunct. We have developed a rule based Bengali *clause boundary detector* to improve copula identification and translation.

The form of the Hindi pronoun depends on the gender and can be obtained by anaphora resolution. The form of the verbs also depends on the gender of the karta. An *anaphora resolution module* is developed to improve the translation of pronouns and verb forms.

A major source of error observed in the Bengali to Hindi rule based machine translation system is incorrect translation of nominal markers. Nominal markers are composed of suffixes and postpositions and the same marker may be translated differently in different contexts. We have developed some rules for translating ambiguous Bengali nominal markers to appropriate Hindi markers. We have also developed rules for correcting some other errors in the baseline system like inability to insert copula verb, wrong boundary and tags of chunks, gender, number and person disagreement between subject and verb, etc.

The baseline system replaces Bengali word with the most frequent Hindi word. But, the most frequent translation may not be the proper translation for a specific context. We propose a technique which finds a better lexical choice among the dictionary options with the help of the contextual information of a Hindi monolingual corpus and a lattice-based data structure.

Even after incorporating the new modules and the transfer grammar rules and modifying some of the existing modules in the Bengali to Hindi MT system, some errors may still remain. The errors we address are spelling error, morphological error, word choice error, insertion error, deletion error, and reordering error. Our idea is to consider different phrases of a given sentence, and find appropriate replacements of some of these from the frequently occurring similar phrases in the monolingual corpus. When looking for similar phrases we consider phrases containing words that are spelling variations of or are similar in meaning to the words in the input phrase. We use a framework where we can consider different ways of splitting a sentence into short phrases and combining them so as to get the best replacement sentence that tries to preserve the meaning meant to be conveyed by the original sentence.

The baseline and modified translation systems are evaluated using the BLEU automatic metric and human evaluation process and latter system is found performing better in both evaluations.