Chapter 21
Word Sense Based Hindi–Tamil Statistical Machine Translation

Vimal Kumar K.
Jaypee Institute of Information Technology, India

Divakar Yadav
Jaypee Institute of Information Technology, India

ABSTRACT

Corpus based natural language processing has emerged with great success in recent years. It is not only used for languages like English, French, Spanish, and Hindi but also is widely used for languages like Tamil, Telugu etc. This paper focuses to increase the accuracy of machine translation from Hindi to Tamil by considering the word's sense as well as its part-of-speech. This system works on word by word translation from Hindi to Tamil language which makes use of additional information such as the preceding words, the current word's part of speech and the word's sense itself. For such a translation system, the frequency of words occurring in the corpus, the tagging of the input words and the probability of the preceding word of the tagged words are required. Wordnet is used to identify various synonym for the words specified in the source language. Among these words, the one which is more relevant to the word specified in source language is considered for the translation to target language. The introduction of the additional information such as part-of-speech tag, preceding word information and semantic analysis has greatly improved the accuracy of the system.

INTRODUCTION

Natural language processing (NLP) is a part of artificial intelligence which interacts with the systems (computer) through natural languages to perform desired actions. It deals with understanding and analyzing human languages in order to perform various functionalities which can enhance the interaction between the machine and the humans. There are various widely used algorithms under NLP, especially statistical natural language processing but each algorithm has its own bottleneck. These algorithms are usually based upon the analysis of large textual corpora and then calculating probabilities in order to

DOI: 10.4018/978-1-7998-0951-7.ch021

achieve the desired results. According to linguistics, corpus refers to large structured texts consisting of numerous words which are used for statistical analysis of the text. Generally, the corpus should be annotated to provide an efficient statistical analysis. The Corpus consists of each and every word in every sentence used for the language analysis. These words are added to the corpus along with the information about its part of speech such as: verbs, adjectives, nouns, and adverbs etc., which are called as POS tags. Corpus based NLP techniques have emerged with great success in the recent years. It is not only used for languages like English, French, Spanish, and Hindi but also is widely used for languages like Tamil, Vietnamese etc.

Numerous algorithms have been introduced in statistical machine translation to provide various intelligent functionalities for human-computer interaction. All these algorithms parse the sentences and then group the words before the translation process. Parsing of free word order languages such as Indian is also a bottleneck in these methods (Bharati et al., 2009; Bharati & Sangal, 1993). Local word grouping (LWG) is basically used in Indian languages since there is a need for grouping the words based on the context in which it is used and the meaning of those words will be clear only when it is grouped together (Bharati et. al, 1991; Ray et. al, 2003; Balaji et al., 2014). In these existing technologies related to corpus based NLP, the statistical analysis for machine translation makes use of a parallel corpus. In a parallel corpus, each word in the source language is mapped parallel with its corresponding word in the target language. In addition to parallel corpus being used for translation, the part-of-speech of the words is also considered for machine translation. Also, in statistical machine translation, the target texts are generated on the basis of statistical models and these models are derived from the analysis of the text corpus of the two languages. Generally, a document is translated to a probable sentence in the target language according to the probability distribution *P(t|h)* which refers to the probability of string *t* in the target language (for example, Tamil) given the string *h* in the source language (for example, Hindi).

Naive Bayes algorithm is one of the existing algorithms which are based on statistical analysis of the existing bilingual corpus. The algorithm uses probability of occurrence of words for translation from one language to another. For a particular word, its probability is calculated based upon the frequency of occurrence of the word in the corpus. The meaning which has maximum occurrence in the target language will be the probable translation for the input word. The mathematical representation of this algorithm is:

$$\mathrm{P}\left(\frac{x}{y}\right) = \mathrm{P}\left(\frac{y}{x}\right) * \mathrm{P}(x) / \mathrm{P}(y)$$

Since P(y) will not affect the result, the equation is equated as shown below:

$$\mathrm{P}\left(\frac{x}{y}\right) \cong \mathrm{P}\left(\frac{y}{x}\right) * \mathrm{P}(x)$$

This mathematical representation signifies that there is translation model and language model being used to perform the translation process. Translation model analyzes the translation between source and target language whereas the language model analyzes the target language being used. These statistical algorithms can further be improved by considering the tag information of the word under consideration but still there will be lack of efficiency and accuracy. Also, ambiguity is one of the concerns of these