Abstract

In the era of globalization, vast amount of information is available on World Wide Web. This information is available in various natural languages. In order to access this information, the speakers of different languages have to rely on translation or learn the other language. Learning multiple languages is not feasible for everyone, hence the need of translation arises. In such a scenario, machine translation emerges as an important discipline in the field of computational linguistics that can give access to the vast information within a short time and at low cost.

Translating natural languages automatically is one of the most complex and comprehensive applications of computational linguistics [63], and the dream of a perfect full-scale machine translation system still remains an intellectually challenging task. Several approaches, such as rule based, statistical and hybrid, have been tried out to build linguistic tools. All these approaches have their own merits as well as demerits.

Pāṇinian Grammar, universally admired for its insightful analysis of Sanskrit [74], offers a perspective to language analysis. The main focus of Pāṇinian Grammar is to explain how, where and how much information a language encodes that helps in capturing information dynamics in a language [3].

In order to use Pāṇinian theory to analyse other languages, it is necessary to model these languages in terms of Pāṇinian primitives. We have selected the primary basic notions such as nominal inflections (*sup*), verbal inflections (*tiṅ*), basic grammatical constructions (*pada*), compound constructions (*samasta-pada*), etc., to analyse the syntax and semantics of natural language sentences. In our research, we have used these concepts to analyse English from Pāṇinian perspective and then map English structures into Hindi for handling three major linguistic tasks for English-Hindi machine translation as follows:

1. **Finding Out the Meaningful Linguistic Units That Can Participate in a Sentence:** This step theoretically accounts for English sentence in terms of Pāṇinian primitives such as *sup*, *tiṅ*, *pada* and *samasta-pada*. Here we find out the syntactic and semantic units called *padas* and *samasta-padas* and the grammatical markings that connect these units, the *padas* and *samasta-padas* with the other meaningful units in a sentence. Such analysis looks at the words in a sentence as a combination of content words and grammatical markings [125]. Thus, this type of syntactic analysis helps in capturing the flow of information and gives us the shallow parse of the sentences. It is also helpful in various other linguistic tasks such as dependency parsing, word ordering and case marking.

2. **Mapping Dependency Parsers' Output into a Uniform Notation:** Parsers play an important role in analyzing the exact meaning of a sentence. The parse tree produced by the parsers formally represents syntactic and semantic information among the words in a sentence. In a machine translation system, various modules like Gender Number Person (GNP) agreement, Word Sense Disambiguation (WSD), reordering of the source language sentence according to target language sentence, etc. are some of the important tasks which require dependency parse as input [89].

   Several dependency parsers are available for English viz. Stanford Parser, Link Grammar Parser, CLEAR Dependency Parser, Minipar, Malt and many more [119, 47, 36, 85]. However, no two dependency parse output formats match with each other. These parsers differ on the names of the dependency labels, their representation style and the number of dependency labels [17].

   Since the parsers are based on different approaches, and each approach has its own limitations, the parsers produce different degrees of output quality for different syntactic constructions. In order to obtain maximum benefit out of these parsers, one has to combine the strengths of the parsers [24]. Hence arises the need of a common representation schema. Also, the use of computers as an information processing device requires a sophisticated theory for processing the information in a language.

   Bringing the parsers' output into one uniform notation provides the system an ability to plug-in any of these parsers without modifying other translation modules, thereby avoiding large amount of manual work that is costly as well as time consuming.

3. **Framing Transfer Grammar Rules:** Word ordering plays an important role in the translation process between languages. Using the concept of *pada* and *samasta-pada*, the source language sentence is manipulated into target language sentence for natural translation. Since the notion of *pada* overtly identifies the grammatical markers and their respective content words as syntactic units of a sentence, we get a little more freedom for word ordering as an advantage due to the relatively free word order nature of the target language Hindi [23].

   The concept of *pada* also helps in generating direct or oblique case for Hindi.

The main focus of this study is on machine translation between two diverse languages: English and Hindi. But the research carried out in this thesis is generic enough to be applied to any English-Indian Language pair. This is because most of the Indian languages have many common features and the transparency of the tools built during this research allows to cover diverse cases easily.