Chapter 1

Introduction

The term Machine Translation (MT) refers to the translation of one natural language text into another natural language text. It is an automatic process, whereby, the information in the natural language, often called Source Language (SL), has to be faithfully transferred into Target Language (TL) with the help of computer softwares.

1.1 Problem Statement

Translating natural languages automatically is considered to be one of the most complex and comprehensive applications of computational linguistics [63]. George Artsrouni, a French-Armenian, received the first patent in July 1933 for ‘translating machines’ and a Russian scientist, Petr Petrovich Troyanskii, in September 1933 [69, 4]. Since then, the research on MT is being carried out all over the world.

The Automated Language Processing Advisory Committee (ALPAC 1966) report, known for stressing the negative aspects in its study, brought to an end the substantial funding of MT research for more than a decade in USA [67]. The report concentrated exclusively on US government and military needs in the analysis and Russian language documents scanning. It was not concerned with other potential uses or users of MT systems or any other languages. Despite that, many government and private institutes as well as individuals have been trying out various approaches for MT.

Since early 1980s, the interest in MT has been growing, but the dream of a perfect full-scale MT system still remains a difficult and intellectually challenging task. This is mainly because natural languages encode linguistic information in a very complex manner. And computer as an information processing device demands a sophisticated theory for processing the information in natural languages.

When it comes to translation between closer languages such as Hindi and Urdu or Hindi and Punjabi etc., the problem of mapping structures is not so difficult. However, as one moves towards translation between linguistically diverse languages such as Hindi and English the problem becomes much more complex. The problem taken up in this thesis is mapping structures such as the basic syntactic units, the dependency parse, and transfer grammar between English and Hindi and other related aspects of translating across these two diverse languages from Pāṇinian perspective.
1.2 Overview of the Solution: Pāṇinian Perspective to Language Analysis and its Application to MT

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

In our attempts, using insights from Pāṇinian Grammar, we try to capture the following three crucial linguistic aspects in MT between two diverse languages namely English and Hindi:

1. Finding Out the Samarthta i.e. the Meaningful Linguistic Units: Pada and Samasta-pada:

   According to Pāṇini, no grammatical operation can take place, be it word formation or sentence formation, unless they qualify the condition of being samarththa. Thus the concept of sāmarthya is a fundamental principle for any grammatical operation in a language string. The words samarthta and sāmarthya are used interchangeably in Sanskrit grammar. There are two types of sāmarthyas:

   (a) Ekārthībhāva Sāmarthya (Single Integrated Meaning): Where the words having direct semantic connection become one word as in compounds and derivational morphology. In this case, the word samarthha means “mingled together” (saṁgaṭṭhaṁ samarthhaṁ) and “fused together” (saṁṣṛṣṭhaṁ samarthhaṁ) [70].

   The objective of ekārthībhāva sāmarthya is to present compounds as one pada (ekapada) or as a single unit. For example, the compositional compounds rāja-puruṣaḥ (king-man) derived from rājāḥ puṣuṣaḥ and yudhīṣṭhiraḥ (one who is always stable in the battle) derived from yudhiṣṭhiraḥ. And also in non-compositional compounds such as kṛṣṇa-saṁarpaḥ (cobra).

   In English, ekārthībhāva sāmarthya can be seen in both compositional as well as non-compositional compounds. For example, ‘lawn tennis’, ‘bird-cage’ and ‘blackbird’.

   (b) Vyapekṣa samarthya (meaning-interdependence): In this case, samarthha means “seen together” (saṁprekṣiṭārthaḥ samarthhaḥ) and “bound together” (saṁbadṛhārthaḥ samarthhaḥ) [70]. For instance, subject, verb, object, etc. are seen mutually semantically bound together in a sentence. The objective of vyapekṣa sāmarthya is to show sentence as a single meaningful unit. Padas seem to carry diverse meanings but a sentence indicates a single meaning.

   The words in a sentence are always linked with each other via one of these sāmarthyas. Hence, the samarthha theory applied to a language would help in identifying the meaningful linguistic units described as padas and samasta-padas in Chapter 3 via ekārthībhāva sāmarthya. We examine how English encodes semantic information through various overt and covert syntactic mechanisms in a sentence and make it explicit for MT purpose.
2. **Mapping Dependency Parse Output into Anusāraka Dependency Schema:** We have developed a dependency schema for English. This schema is based on Pāṇinian way of sentence analysis. In order to mark semantic relations among the words in a sentence, we use syntactic cues from Pāṇinian concepts such as *sup, tin̄, pada* and *vyapekṣā sāmarthya*. Here, we explore different mechanisms. We map the output of different dependency parsers into our framework for standardization. Once the output of different dependency parsers is brought into one standard format, one can plug-in these parsers into own MT system without any changes in subsequent modular organization. Thus, it helps in building a robust MT system and brings down the number of dependency labels.

3. **Framing Transfer Grammar Rules:** Word ordering plays an important role in machine translation. Using the the concept of *pada* and *samasta-pada*, the source language (SL) sentence is manipulated into target language (TL) sentence. Since the notion of *pada* and *samasta-pada* explicitly identifies the overt and covert syntactic devices that bind the words together in a sentence, reordering a source language sentence into a target language sentence becomes easy, especially for languages which are morphologically rich and have relatively free word order such as Hindi and other Indian languages.

The notion of *pada* is also useful with respect to word form generation and handling of direct and oblique cases in Hindi.

Pāṇinian way of analyzing a language draws generalizations whereby it brings down the number of rules subsequently. The tools that we have developed using Pāṇinian insights are essentially built for English-Hindi MT. But these tools are generic enough to apply to any English-Indian language pair with some modifications.

### 1.3 Contributions

We show how concepts from Pāṇinian grammar can be applied in MT for faithful transference of the source language information into the target language. Using insights from Pāṇinian grammar, we have built tools that accomplish the following major linguistic tasks:

- **Identification of Padas (Meaningful Linguistic Units):** This tool identifies the meaningful linguistic units called *padas*. A *pada* is a syntactic unit, a unit to which its semantic connection with respect to other words in a sentence has been marked explicitly.

We have taken insights from Pāṇini’s grammar to identify various types of nominal and verbal inflections for English. These insights are then used to identify the simple as well as complex syntactic units in English.
• **Developing a Standard Dependency Schema for Indian Languages:** Dependency parse is preferred over constituency parse [85, 48] because it suits a wide range of natural language processing tasks viz. as question answering, information extraction, machine translation, etc. [46].

The research is motivated from arriving at a common representation for English to Indian languages, especially English to Hindi MT system Anusāraka that uses dependency parse for analysis of the source language sentences. Various translation sub-tasks such as Gender Number Person (GNP) agreement, Word Sense Disambiguation (WSD), reordering of the source language sentence according to the target language sentence, etc., require dependency parse labels as input [89]. However, different parsers have different output schemes which differ in relation labels as well the number of the dependency labels. Bringing parsers’ output into one uniform notation provides the system an ability to plug-in any of these parsers without modifying the other subsequent translation modules, thereby, avoiding large amount of manual work which is costly as well as time consuming.

• **Developing a Transfer Grammar Tool:** Hindi word order differs from English. That is why reordering of source language words according to target language word order plays a crucial role. We have used the concept of *pada* that groups the content words with the function words and places them according to the target language word order. And the reordering rules based on “mirror structure” along with some exceptions arrange the *padas* into a fluent target language word order.

The main focus of this study is on English-Hindi language pair. But we have claimed that the research carried out in this thesis is generic enough to be applied to any English-Indian language pair. Because most of these languages have many common features and the transparency of the tools built during this research allows to cover diverse cases easily.