Chapter 1

Overview

Machine Translation (MT) is among one of the first applications of computers. Since 1950, there have been several efforts [44] in this field leading to the development of several products and tools for language analysis. With the emergence of World Wide Web, enormous amount of corpora is being available which has boosted the research in Statistical techniques in the field of Natural Language Processing (NLP). The emergence of new branches of knowledge such as computational linguistics (CL), advancements in Statistical Techniques, developments of computational grammars such as Lexical Functional Grammar (LFG), Tree Adjoint Grammar (TAG), development of parsers based on these and various other formalisms such as minimalism, Link grammar, HPSG, etc., and the availability of corpora for several languages are the key factors for the growing activities in the field of NLP in general and MT in particular.

The conventional machine translation systems are fragile and do not provide a fall-back mechanism. Since the main focus of these systems is also to produce a translation, ‘faithfulness’ to the original text takes a back seat. With an aim to provide a faithful access to the text in other languages, anusaaraka [70] was developed. With appropriate division of load between man and machine, Kannada-Hindi anusaaraka...
demonstrated that it is possible to reduce the language barrier. It was necessary for an anusaaraka reader to undergo some training on the syntactic divergences and special notation used to handle the semantic divergences between the source and the target language. Following the success of Kannada - Hindi anusaaraka system, development of 4 other anusaaraka systems among Indian Languages was undertaken jointly by IIT Kanpur and the University of Hyderabad in the early 90’s. Unavailability of electronic texts in Indian languages necessary for building certain Statistical tools for language analysis, and the tremendous amount of efforts required to develop such tools manually using rule based approach put a check on further development of the anusaaraka.

With an increasing interest in MT, several efforts were undertaken by various research groups to develop language resources and tools for language analysis. Several of these resources and tools for English are available freely. With the growing need of the Indian society to have access to the English texts, and the availability of various English language resources and analysis tools, the anusaaraka group shifted its focus from building anusaarakas among Indian languages to building anusaaraka systems from English to Indian languages (with Hindi as a case study). Though various analysis tools for English were available, one can not just plug-in these components in a MT system, since these resources follow different grammar formalisms and thus do not produce an uniform output.

Thus there are two requirements – one is to provide faithful translation with reduced burden on the user and the second one is a facility to plug-in available resources thereby avoiding the re-inventing of the wheel. The goal of this thesis is to present an architecture for MT system that produces a robust and faithful output, allows plugging-in of different resources, degrades gracefully in case of failures, provides an interface to display right kind of information at right time and provides full control to the user who can navigate through various linguistic analyses and resources as per
the need.

The earlier amusaaraka system could be used only by those who have undergone some training. This puts a considerable load on the part of user. When the divergence between source and the target language is more (e.g. as in the case of English and Hindi), the load on the user is substantial. The proposed architecture reduces this burden. This architecture is planned to cater to the needs of diverse requirements such as fidelity of the translated output and naturalness of the translation.

On the face of it, this proposal may seem to be ambitious, but the insights from Indian Grammatical Tradition (IGT) and in particular the information centric analysis approach of Pāṇini provides appropriate clues leading to the required architecture.

“Pāṇini’s grammar is universally admired for its insightful analysis of Sanskrit” [56]. In spite of being a grammar basically written for Sanskrit, it provides many ingenious concepts for language analysis, which are universal in nature. Pāṇinian Grammar (PG), as any other grammar formalism, provides an appropriate set of procedures to identify the relations among words in a sentence. However, the importance of PG lies in the minute observations of Pāṇini regarding the information coding in a language.

A concept of śabdasūtra which is influenced by the notion of pravṛtti-nimitta guarantees faithfulness. Śabdasūtra provides a core sense (or in some sense a nuclear sense) of a word. The use of pravṛtti-nimitta to express the meaning of English into Hindi together with the other factors such as ākāṅkṣā(expectancy), yogyatā(competancy), tātparya(intention) and sannidhi(proximity) help in the process of śabdabodha (understanding the meaning of a sentence).

These four factors together with the arthабodhакatva or gamакatva (ability to convey
Chapter 1: Overview

the desired meaning) and the clues from information centric analysis by Pāṇini provided guidelines for studying the structural divergences between English and Hindi. In case machine fails to provide translation, with the help of the knowledge of these divergences and the Śabdasyūtras a user can still ‘understand’ the original English sentence ensuring fidelity.

The parallel processing of various modules and the changes in the order of operations guarantee the robustness and graceful degradation in case of failures. The concept of interfaces to parsers provides a plugging-in facility. A voting algorithm allows to plug-in more than one parser (or tagger) facilitating the selection of the best parse (or tag).

Thus this thesis consists of two parts – one part deals with the engineering aspect or the design of a system while the other part provides a sound scientific base for the design, ensuring that the design is based on the time tested principles thereby ruling out any possibility of ad-hoc solution.

Chapter two provides a brief history of the MT efforts within and outside India and also gives a brief summary of the new trends in this field. Typically the problems in MT arise because of the differences in two languages at various linguistic components such as syntactic, semantic, pragmatic etc. It is natural that the literature on MT discusses these problems in the light of these linguistic modules. Chapter three on the other hand, presents the problems in MT from information centric point of view. In this chapter, the process of MT is viewed as a process of decoding and encoding. Since languages code information only partially and sometimes on discontinuous strings even at far off places, decoding a string in a language can not yield the ‘complete’ picture. Further at the encoding point, if the two languages differ in the convention of coding, the resulting encoding may lead to catastrophe while dealing with the mapping the information in SL to TL with respect to the differences at the
level of labeling and packaging of information may crop up. In the fourth chapter, we look at the problem of MT afresh, and provide a new architecture with the desired features. We mainly discuss the proposed architecture for English - Hindi language pair. We claim that this architecture provides a better approach for MT because it is robust and transparent. Since it is based on the ‘information dynamics’, it also helps a developer in discovering the divergence cases more easily. Further the architecture is flexible enough to use both rule based as well as statistical modules allowing us to take an eclectic approach.

In the second part we show the appropriateness of Pāṇini’s theory (chapter 5) to provide a scientific base for carrying out the information centric analysis. This study further leads to an important observation regarding the limitations of MT. The questions viz. where does a language code information?, how much information does it code?, and the manner in which it codes the information are the three aspects of the information dynamics or the parameters that are crucial in identifying the “true nature of the language”. These three parameters may be used to determine the syntactic divergence between the languages. And hence we claim that any grammar which is developed with the three questions in mind: where, how much and how is the information coded, would be truly in Pāṇinian spirit. The insight obtained from Pāṇini’s work is used to discover the reasons behind the structural divergence between English and Hindi. Chapter 6 discusses these structural divergences and concludes that all these structural divergences may be attributed to only the two missing formatives of the accusative and the yes-no question in English. It further lists three important things a Hindi reader reading an English text should tune to. They are

a) acquire a new ‘vr̥tti’ – the ‘quazi compound’ \( V_n \)

b) do away with the normal ‘sannidhi’ (proximity) between a verb and its auxiliary and also between a noun and its post-position (which are integral part of Indian lan-
guages), and acquire new ‘sannidhi’s between: i) a subject and auxiliary and ii) a verb and its preposition, and

(c) remember that the occupant of subject position need not have any kāraka role with the corresponding verb.

Seventh chapter explains in brief, how Dorr’s divergences are handled at various levels of anusaaraka output and then mentions the three important syntactic phenomena in English viz. a) resultative constructions, b) verbs of motion specifying the manner of motion, and c) the absolutive constructions. As a consequence of these structures being absent in Hindi, there is a structural gap between English and Hindi. Hence such constructions are really problematic. Eighth chapter discusses the concept of pravrtti-nimitta and how it helps in developing śabdasūtras for words with various types of lexical divergences. Finally with the help of examples, we explain, how the meaning of an English sentence is understood by the user following the anusaaraka outputs at various levels leading to the Śābdabodha. Ninth chapter discusses the issues involved in the development of parser interfaces. The current trend is, most of the parsers produce both a phrase structure as well as dependency style output. But the output of various dependency styles do not match. We propose guidelines for producing the output in dependency style based on Pāñinian Grammar Formalism. The tenth chapter provides future directions for taking anusaaraka further.

The central theme of the thesis is the information centric analysis. An effort is made to run this thread of information centric analysis through out the thesis.