INTRODUCTION

It was my reading of the article ‘bhaashaantara mimaansaa’ (criticism of translation) in B. Nemade’s book ‘saahityaachii bhaashaa’ (language of literature) that led me to enter into the field of ‘theory of translation’. In due course, I studied linguistics and thereafter computational linguistics. While I was working on various projects such as Marathi wordnet, Marathi Spell Checker, Marathi POS tagger and UNL at I. I. T. Mumbai, I was drawn to the complex subject of Machine translation. These events together took me to the path of research in word sense disambiguation. My guru Dr. P. P. Giridhar provided enlightenment when I started my study in the field of word sense disambiguation. As I am comfortable with the language triplet of Marathi, Hindi and English being resident of Maharashtra and citizen of India, I decided to work on this language triplet.

Being Indian languages, Marathi and Hindi are rich in morphology but poor in resources. English being a language of developed countries which is understood the world over, loads of resources are available even on internet and that also free of cost. The world wide resources for natural language processing are much advanced for many languages. The research in the field of machine translation in general and word sense disambiguation in particular is latterly heading towards statistical methods. But unfortunately, in case of Marathi and Hindi, due to scarcity of resources in general and inadequate corpus in particular, statistical applications are yet to be developed. Anyhow, even for using statistical methods one needs rule base to some extent. It is on this background that the current work is concentrated on word sense disambiguation for translation from Marathi to Hindi and English. Here we use the expression word sense disambiguation to mean resolving ambiguity at word level.

At I. I. T. Mumbai, theoretical background of Marathi POS tagger is well founded. As a part of this study, theoretical foundation for resolving ambiguity at part of speech level is also developed. This facilitates to assume that the ambiguity at part of speech level can be resolved by assigning tags to the words in a text with respective parts of speech. The present work is focused on polysemous words. Further, while doing so the situation of translation from Marathi to Hindi and English is considered and not otherwise.

Though the scope of the work is thus narrowed, the methodology is applied and research is designed in such a way that the same can be used for any number of languages and in any direction of translation.

This is a pioneer work in the scene of Indian languages. The present researcher had worked right from the ground level. Besides the heavy readings in the field, several practical exercises were also undertaken.
In order to understand the theoretical background of the process of translation a comparative study of a well known Marathi translation and the original English text was undertaken. This was coupled with an interview of the translator in person. The real world experience is taken into account by interviewing Marathi native speakers to ascertain how they resolve ambiguity. To understand issues involved in word sense disambiguation some multilingual persons were examined as a part of field work. Results are listed after analyzing the questionnaires filled by them which provided insights to the issue.

The Marathi corpus developed at CIIL, Mysore, proved to be of use in concluding meanings of words in context. This corpus was found to be inadequate in some cases while searching key words pointing to a particular meaning of a word in its specific usage. Such situations are overcome by accessing the electronic versions of Marathi news papers or Marathi wordnet.

Since no parallel corpora in case of the language triplet of Marathi, Hindi and English are available, possible translations of a given Marathi word into Hindi and English were confirmed with the help of dictionaries and heuristics.

The present researcher chose to handle cases of words belonging to all parts of speech with the help of the context of single sentence rather than concentrating on words belonging to single or few parts of speech with the help of the context of required size needed for disambiguation. This is because it is found that at times when context of single sentence is inadequate for word sense disambiguation; key words pointing exact meaning of the given word may be outside the entire discourse.

While reading extensively in the field, it is observed that the significance of ambiguity is completely different in case of humans and machines. Machine need to deal with entities which have potential for ambiguous readings that tend to be completely unnoticed in human discourse / conversation. One may miss even a pun, an instance of real ambiguity. Words may be polysemous in principle, but remain unrealized and still understood by a person. Machine need to resolve ambiguity at every phase including several interpretations of affixes. Even disambiguation of morphological units is a very complex process for a machine.

This thesis covers cases of words belonging to all parts of speech considering the subclasses of each of them. Further, it describes minutely the procedures to disambiguate the several interpretations of the Marathi verb suffix ‘to’. It is a pleasure to submit a CD that demonstrates the software developed for disambiguation of suffix ‘to’. The task of disambiguation which was completed satisfactorily is explained in the sixth chapter of this thesis.

Chapter one discusses the comparison between man and machine on the line of cognitive science in general and artificial intelligence in particular. The chapter on
translation as phenomenon (chapter 2) deals with theoretical background of translation and is supported by some practical work. The nature of ambiguity (chapter 3) is sought to be understood by knowing its types and levels. The principles of determining meaning of a linguistic entity discussed in ‘waakyapadiya’ that is in Sanskrit tradition throw light on how context points to a particular meaning of a given word in its given usage. Chapter four discusses the ambiguity inherent in the word ‘word’ and the relation between the word and its sense in terms of the concept it refers to. It further describes the previous worldwide efforts and the tools developed in the field of word sense disambiguation. The discussion on category wise word sense disambiguation is presented in chapter five. It lists the key words that help in resolving ambiguity in a given situation. For the purpose of this task, the different senses of word could be fixed by considering only those distinctions that are lexicalized cross linguistically. Chapter six includes the description of procedure to disambiguate Marathi verb suffix ‘to’. It also describes the way in which the software for disambiguation of suffix ‘to’ is developed and the evaluation of that software.

We are now confident that it is possible to complete the task of word sense disambiguation with the help of textual information. The real world knowledge can be represented in the form of machine readable data. With the help of these data machine can perform like a human at least in the field of word sense disambiguation.

Looking from this work to the future, we see a path to completely automated machine translation. We are aware that this aim can be achieved only after availability of a full range of workable NLP (natural language processing) tools like spell checkers, POS taggers, user specific search engines, chunkers, sense taggers etc. But, comprehensive monolingual corpus as well as multilingual corpora are a must before moving further in this field.