Machine translation is about automatic translation of one natural language text to another using computer. In this thesis, morphology based Factored Statistical Machine Translation system (F-SMT) is proposed for translating sentence from English to Tamil. Tamil linguistic tools such as Part-of-Speech Tagger, Morphological Analyzer and Morphological Generator are also developed as a part of this research work. Conventionally, rule-based approaches are employed for developing Machine Translation. It uses transfer-rules between the source language and the target language for producing grammatical translations. The major drawback of this approach is that it always requires the help of a good linguist for the rule improvement. So, recently data-driven approaches such as example-based and statistical based systems are getting more attention from research community. Currently, Statistical Machine Translation (SMT) systems are playing a major role in developing translation between languages. The main advantage of using Statistical Machine Translation system is that it is language independent and it disambiguates the sense automatically with the use of large quantities of parallel corpora. SMT system considers the translation problem as a machine learning problem.

Statistical learning methods perform translation based on large amounts of parallel training data. At first, non-structural information and statistical parameters are derived from the bi-lingual corpora. These statistical parameters are then used for translation. Baseline Statistical Machine Translation system considers only surface forms and does not use linguistic knowledge of the languages. Therefore its performance is better for similar language pair when compared to the dissimilar language pair. Translating English into morphologically rich languages is a challenging task. Because of the highly rich morphological nature of Tamil language, a simple lexical mapping alone does not help for retrieving and mapping all the morphological and syntactic information from the English language sentences.

Tamil word forms are productive, that is, word forms are written without spaces. Inflected forms of Tamil words are separate words in Tamil. This leads to the problem of sparse data. It is very difficult to collect or create a parallel corpus which contains all the possible Tamil surface words. Because, a single Tamil root verb is
inflected into more than ten thousand different forms. Moreover, selecting a correct Tamil word or phrase during translation is a challenging job. The corpus size and quality decides the accuracy of the Machine Translation system. The limited availability of parallel corpora for English-Tamil language and high inflectional variation increases the data sparseness problem for baseline phrase-based SMT system. While translating from English to Tamil language, the SMT baseline system will not generate the Tamil word forms that are not present in the training corpora.

The proposed Machine Translation system is based on factored Statistical Machine Translation models. The words are factored into lemma and inflected forms based on their part of speech. This factorization reduces the data sparseness in decoding. Factored translation models allow the integration of the linguistic information into a phrase-based translation model. These linguistic features are treated as separate tokens during the factored training process. Baseline SMT system uses untagged corpora for training, whereas factored SMT uses linguistically factored corpora. Pre-processing phase allows including language specific knowledge into the parallel corpus indirectly. In preprocessing, bi-lingual corpora are converted into factored bi-lingual corpora using linguistic tools and reordering rules. Similarly, Tamil language sentences are also pre-processed using the proposed linguistic tools like POS tagger and Morphological analyzer. These factored corpora are then given to the Statistical Machine Translation models for training. Finally, Tamil morphological generator is used for generating a surface word from output factors.