CHAPTER 9

CONCLUSION AND SCOPE FOR FUTURE WORK

9.1 CONCLUSION

The investigations carried out in this research focus on the improvement of language and parsing models in Tamil language. Language model adaptation using CLSA has been carried out. Initially, domain independent model has been built and adapted directly with translation lexicons. This approach has boosted the probabilities of content words. Owing to the application of partial morphology, morphologically related Tamil words are also boosted up. Subsequently, topic specific models have been built and used for adaptation in the domain independent model after topic identification. This approach has further enhanced the probabilities of content words in the domain independent model. Significant improvements in perplexity have been observed in topic specific and domain independent models. WER has been obtained with reasonable improvement from ASR with initial and adapted domain independent models.

For the generation of the large annotated Treebank, POS, phrasal and dependency tag sets for Tamil have been created with more than 600, 11 and 31 tags respectively. Initially, 3,261 sentences have been used for the generation of phrase and dependency structure Treebanks through hand annotation and bootstrapping. Then the rule based morphological analyzer and POS tagger, and phrasing tool for Tamil have been developed. These tools have been tested with their respective gold standards and a better performance has been achieved. POS tagged Tamil sentences have been
obtained as a resource through statistical alignment and projection technique. Similarly, root words have been obtained through statistical alignment, lemmatization and morphological induction. The rule based morphological analyzer and POS tagger have been improved with stems and categorical information obtained from morphological induction and POS projection respectively.

Using this improved rule based approach, phrase and dependency structure Treebanks with 10,000 sentences have been developed. Reliable phrase and dependency parsing models have been built using these Treebanks, tested with a test case of 1,000 sentences and evaluated against gold standards. Phrase structure model has given good accuracy and covered the challenging features to a certain extent. Non-projective dependency parsing model has given better accuracy than phrase structure model and it covered the same to a greater extent. As a result, significant contributions have been made in the improvement of language and parsing models in Tamil language.

9.2 SCOPE FOR FUTURE WORK

In the improvement of language model, present approach adapted unigram translation probabilities with trigram language model. In the future, similarity in phrasal or chunking structures can be studied. The noun with cases in Tamil is equivalent to the noun with prepositions in English. Verb groups in Tamil are equivalent to Verb and its auxiliaries in English. Prefix with noun in Tamil is equivalent to determiner with noun in English. Similar chunks or phrases can be identified. These cases may be treated as single or separate entities in English and Tamil. Without partial morphology, CLSA approach may be applied to study N-gram possibilities in English documents and higher order adaptations may be done with bigram and trigram probabilities in Tamil language trigram model. For the topic specific models,
more and more documents related to trained documents can be projected into reduced CLSA space. For totally diverse documents, CLSA space can be broadened to encompass many more topics. Similarly, the alignment model can be modified to accommodate similarities of English and Tamil phrases.

In the improvement of parsing models, more morphological features can be applied in the form of POS tagging by the generation and application of new tags which are more specialized and unique in nature. Besides these features, more phrasal tags and rules can be added. For dependency annotation, semi-automated or automated tools can be developed. The dynamic memory learning for partial stems can be refined to reduce partial stems for the same stem. Likewise, context based disambiguation techniques can be used in improved rule based approach. Some of the issues in Tamil to be resolved are disambiguation of participle noun and verb, similar adverb and preposition suffixes, adjective and noun clashes and similar suffixes for cases. Lastly, multiple prepositional suffixes added in a direct preposition, wh-words with prepositional suffixes and handling of prefix and circumfix in a word also can be resolved.