CHAPTER 9
SUMMARY AND CONCLUSION

A summary of the work reported in this thesis is presented in this chapter. The conclusion, which follows after the summary, attempts to highlight the contribution to the field of NLP in Kannada language. At the same time, the limitations of the developed systems are also pointed out, so that researchers who are interested in extending any of this work can easily explore the possibilities.

9.1 SUMMARY OF WORK DONE

In this section, the work done is summarized with reference to the objectives of the proposed work. This thesis mainly highlights five different developments in the field of Kannada language NLP which are listed below. The detailed summary of the individual work is given at the end of each chapter.

- English to Kannada Machine Transliteration: Developed using two different approaches. These developments are summarized at the end of the chapter 4.

- Parts of Speech Tagging for Kannada: A statistical based POS tagger for Kannada language is developed using machine learning approach. The proposed development is summarized at the end of the chapter 5.

- Morphological Analyzer and Generator for Kannada: Using FST, a rule based MAG for Kannada language is developed. A statistical based Morphological Analyzer for Kannada verbs is also developed using SVM. These developments are summarized at the end of the chapter 6.

- Syntactic Parse for Kannada: A statistical Penn Tree bank based Syntactic Parser for Kannada language is developed as a part of the research work. This is just a prototype parser model based on only 1000 divorced sentences. The proposed development is summarized at the end of the chapter 7.

- English to Kannada MT: Developed a rule based MT system for English to Kannada language by integrating different modules and various computational linguistic tools.
The MT system successfully works for almost all types of Simple, Continues, Perfect, Perfect Continues sentences and their Negatives. The system also successfully translates the above said sentences with can, may, should, must and their negatives.

This part of the proposed research work is summarized at the end of chapter 8.

The objectives of the work is mapped to the publications which have resulted from this work is as shown in Table 9.1. A detailed literature survey on NLP for Kannada language as well as Indian languages were performed as a part of the proposed research work. The outcome of this literature survey is also mapped to the publications.

Table 9.1: Mapping of Outcomes of the Research to the Publications.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development of an English to Kannada machine transliteration system for translating named entities from English to Kannada language.</td>
<td>Feature Extraction Based English to Kannada Transliteration, Third International Conference on Semantic E-business and Enterprise Computing (SEEC-2010).</td>
</tr>
<tr>
<td>Development of an English to Kannada MT system for performing English language to Kannada language translation.</td>
<td>A Rule Based Machine Translation System for English to Kannada Language”, Journal of Computer Science (Accepted), Published by Science Publication, Indexed in SCOPUS.</td>
</tr>
<tr>
<td>Literature survey on MT for Indian languages.</td>
<td>MT Approaches and Survey for Indian Languages, International Journal of Computational Linguistics and Chinese Language Processing (Accepted).</td>
</tr>
</tbody>
</table>

9.2 CONCLUSION

The major achievement of this work has been the development of a rule based English to Kannada MT system by integrating different modules and source and target language computational linguistic tools. A full fledged rule based MAG tool using FST and a
statistical based English to Kannada machine transliteration systems have been developed as a part of this MT system. The system also uses the well known Stanford PCFG English parser for assigning the proper POS categories to the words in a sentence as well as to obtain the syntactic structure of the English sentence. The performance of the proposed MT depends on the performance of each and every modules and linguistic tools used in the system.

A rule based English to Kannada machine transliteration system also developed using WEKA’s C4.5 decision tree algorithm. Even though the rule based transliteration method consist more complex procedure than the statistical method, it resulted in lower performance than the statistical SVM based transliteration system.

A statistical based morphological analyzer for Kannada verbs has been developed using sequence labelling approach. The well known SVM supervised machine learning algorithm is used to implement the system. The corpus creation was the most time consuming work in this part of research work. Even though the proposed system can handle analyses of complex words, it is specific to the verb category.

This thesis also describes about the development of a POS tagger and a Syntactic Parser for Kannada language as a part of the proposed research work. Both of these developed tools are based on statistical approach implemented using the SVM algorithm. The syntactic parser system uses the developed POS tagger for assigning grammatical category to each and every word in the training and test sentences. These tools can be used in future to implement a MT system between Kannada to any other language, especially Dravidian languages like Telugu, Malayalam and Tamil.

Developing computational linguistic tools like Machine Transliterator, POS Tagger, MAG, Syntactic Parser and MT system by considering all types of peculiarities of the language are challenging and demanding tasks especially for highly agglutinative language like Kannada. The proposed work aims at incorporating more lexical information of Kannada and generates NLP models with good semantic features, which solves the problem more effectively. The performance of the statistical system mainly depends on the size and correctness of the aligned bilingual corpus. If the corpus consist of all types of word categories, then it is possible for a learning algorithm to extract all required features associated with a particular n-gram. On the other hand, the performance of the rule based
system depends on the number and complexity of computational linguistic rules, which is able to cover all types of word and sentence forms. The applications of the proposed linguistic tools can also be used in other NLP tasks such as IR, cross-language applications, data mining, IE, question answering and word sense disambiguation etc.

The non-availability of large scale electronic data resources and inherent complexities of the language are the two main hurdles that have to be solved effectively in order to achieve reasonable developments in Kannada NLP. The success of any NLP developments mainly depends on the coordination and combined work between the proficient linguist and researcher. The lack of linguistic experts in Kannada language is another reason for the present immature stage of the Kannada NLP. The available resources are not directly usable for computationally developing most of the linguistic tools. Because of all these reasons, researcher hesitates to initiate research in Kannada NLP.

As indicated in the literature survey, so far, there are no published works for English to Kannada transliterator, POS tagger and Syntactic Parser systems for Kannada language. Even though there were attempts to develop MAG’s for Kannada and English to Kannada MT system, they are not publically available for comparison with the proposed systems. More over the proposed MT and MAG systems presented in this thesis are the first rule based developments in Kannada NLP. The overall goal of this research work was to develop a rule based MT as well as computational linguistic tools, as at present, Kannada language is in its immature stage in Kannada NLP scenario. That goal, it is felt, has been achieved.

9.3 LIMITATIONS AND FURTHER SCOPE OF THIS WORK

As mentioned in the literature, work in Kannada NLP is in its initial stage and this research was carried out to address a certain problem in Kannada language NLP. All of these developed computational linguistic tools and MT systems are prototype versions but scalable, so that researchers who are interested in extending any of this work can easily explore the possibilities.

The developed English to Kannada machine transliteration systems were based on feature extracted from a parallel corpus of forty thousand Indian place names. The performance of the proposed systems were better for translating Indian place names from
English to Kannada as compared to other named entities like person names, technical terms, foreign names etc. From the experiments it is found that performance of the proposed system increases with increase in corpus size. A complete transliteration system can be developed with a well organized corpus of all types of named entities.

Similarly, the proposed POS tagger was developed with a corpus size of fifty four thousand words taken from the Kannada news paper “ThatsKannada”. The performance of this system can be further improved by increasing the corpus size with more words from diverse sentences.

Even though the developed rule based MAG is a generic system, it cannot handle complex words at this stage. This drawback can be solved in future by writing orthographic and sandhi rules for complex word categories. On the other hand, the developed statistical based morphological analyzer is domain specific to Kannada verb words, but it can handle complex verb words effectively.

The developed statistical based syntactic parser is just a prototype model with a small sized corpus consisting thousand sentences. A fully fledged Kannada syntactic parser can be developed with a large sized well organized Penn Tree bank Kannada corpus.

In addition to the size of the lexicon database as well as the number and diversity of the rules in the “rule file” module, the performance of the proposed rule based MT system also depends on the performance of each and every linguistic tool that are used to develop the system. As in any MT system, Word Sense Disambiguation is one of the major problem and drawback of the developed system that is yet to be solved.