Chapter 7

CONCLUSIONS & FUTURE WORK

7.1 Conclusions

In this work, we have presented an effective POS tagger for Nepali language using hybrid based approach .i.e., Hidden Markov Model (HMM) integrated with Rule-Based method. We have worked with methods so that small amount of tagged resources can be used to effectively carry out the part of speech tagging task. The present work investigated the effectiveness of various POS tagging algorithms and computational linguistic approaches from a natural language processing perspective.

The present study began with the obtainment of a general overview of the part of speech tagging and its different paradigms and approaches. In Chapter 3, a study has been done on some prominent tagging approaches like HMM, CRF model, Rule based approach on their applications in part of speech tagging. It is learned that statistical techniques were more successful than rule based methods, but there are many constraints to adopt the statistical methods for POS tagging. A large tagged corpus is required to develop a statistical model and it gives the output with high accuracy rate.

In Chapter 4, we have described Nepali POS tagset. The size of our tagged corpus is more than 20,000 words and the number of tags in our tagset is 43. We have tested this tagger in four different test corpora, where size of each corpus is about 1000 words. For each test set we have got accuracy in the range of 84 to 93 percent. It is obvious that accuracy will be higher when all the input words feature in the pre-tagged corpus. Since unknown word handling is the major problem in constructing an efficient tagger, we believe when the size of the pre-tagged corpus increases, the accuracy of our tagger would also increase.

In this dissertation, the average overall accuracy of this tagger for morphologically rich and order free language ‑Nepali is 93.50%. Our hybrid models of POS tagger have a much higher accuracy than any other model of Indian and Nepali language. However, the performance of the current model is not as good as that of the contemporary POS-taggers available for English and other European languages. We believe that further error analysis, increasing the size of tagged corpus and more language specific features would improve the system performance.

7.2 Future Work

Part of speech tagger implemented above gives an accuracy of 93.50%. An obvious extension is to improve the accuracy up to 99.9%. A good future work is to analyze the implemented statistical technique and add heuristics to help the tagger in disambiguating the tags and improve the accuracy. For our project 20,000 words were used as lexicon for the tagger. For future work, larger lexicon can also be built which will significantly improve the accuracy of the tagger.

In this paper, we have presented an approach for Part of Speech Tagger of Nepali Language using HMM with Rule Based Approach. It is completely possible to increase a probability-based tagger’s accuracy by applying the contextual transformation rules, since there are always linguistic patterns that cannot be captured by probability-based methods. However, to do so, the tagger must know the tags of the contextual words. The more accurate the surrounding words' tags are, the better result will be produced by applying these contextual constraints.

In this project, there is another problem in handling the words that can act as both common noun and proper noun. So it becomes difficult for the system to tag the word correctly. When such a situation occurs, system tags the word as a common noun. There is high probability that word will be a common noun but in few cases it can act as proper noun. This limitation can be handled by using Nepali Named Entity Recognition system in future.