Chapter V

Conclusion

5.0 Conclusion

In this work, we have exposed the research carried out on applying statistical and machine learning based algorithm to the POS tagging problem. We have used machine learning approaches to develop a part of speech tagger for Kashmiri. However, no tagged corpus was available to us for use in this task. We had to start with creating tagged resources for Kashmiri. Manual part of speech tagging is quite a time consuming and difficult process. So we have worked with methods so that small amount of tagged resources can be used to effectively carry on the part of speech tagging task. We have developed around 2,00,000 word annotated corpora for Kashmiri that has been used for the experiments.

We have used Hidden Markov Model to acquire statistical knowledge about part-of-speech ambiguities for the use in disambiguation algorithm. The HMM models described in this thesis are very simple and efficient for automatic tagging even when the amount of available labeled text is small. The best performance is achieved for the supervised bigram HMM learning model along with morphological restriction on the possible grammatical categories of a word and suffix information for handling unknown words. Although HMM performs reasonably well for part-of-speech disambiguation task but, it uses local features (*current word, previous one or two tags*) for POS tagging.

Part of speech tagger development done in the present thesis was basically a four task effort.

i. **Corpus collection:** This involved building up of an e-Corpus which could be used for the thesis. This was even more important considering that there is no available e-Corpus in Kashmiri till date.The corpus was developed by manually typing in the data using different texts having different themes/domains.

ii. **Tagset development:** This stage involved building up a representative hierarchical tagset which was again a first time effort.The tagset was developed by following the Eagles guidelines. The Eagles guidelines outline a set of features for tagsets, in which some are obligatory, some recommended, and some optional. The obligatory feature of an Eagles compatible tagset is a set of major word categories. The recommended and optional attributes are then organized by these major word categories. The overall number of Category Tags used in the proposed tagset is 26, Type tags are 21 with their corresponding attributes.The reason for developing a hierarchical tagset was to build a tagset which would be useful for many other purposes like Chunking, Parsing, dictionary development, and moreover the present tagset will be very useful for the development of morph analyzer.

iii. **Annotation:** The corpus developed was then manually tagged using the proposed Tagset (hierarchical).

iv. **Development of a Tagger:** Developing a POS tagger is the first attempt towards building a natural language processing (NLP) tool for Kashmiri.

Once the data was tagged it was used for training the tagger (TnT) using different. Tagger was developed for assigning a tag to each word in the corpus, implementing the tagset and tagging-scheme in a tag-assignment algorithm. In other words we can say the output of the tagger consists of POS-tagged files, containing all possible tags for each token, together with the probability of that tag. Accuracy of the present tagger is about 97.64% which would be increased by retraining the tagger by using more and more corrected data.

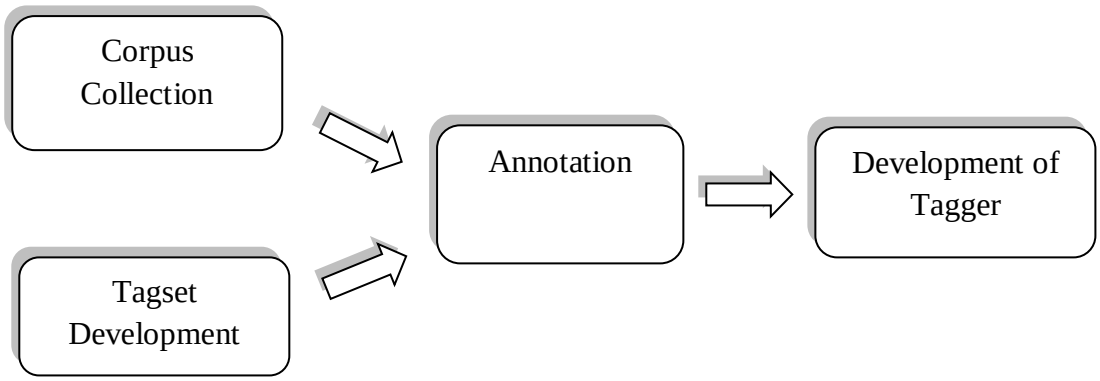

Fig 5.1: Schematic diagram of Tagging process.

The presentwork is a preliminary step in Kashmiri NLP and should have the path for more work in this direction.

■■■