CHAPTER 2

Review of Literature

The area of part of speech tagging using computational techniques has been enriched over the last few decades by contribution from several researchers. Since its inception in the middle sixties and seventies [36], [49], [31], many new concepts have been introduced to improve the efficiency of the tagger and to develop the part of speech taggers for different languages. Initially, people manually engineered rules for tagging. Linguistic taggers incorporate the knowledge as a set of rules or constraints written by linguists. More recently several statistical or probabilistic models have been used for the POS tagging task for providing transportable adaptive taggers. Several sophisticated machine learning algorithms have been developed that acquire more robust information. In general all the statistical models rely on manually tagged corpora to learn the underling language model, which is difficult to acquire for a new language. Hence, some of the recent works focus on semi-supervised and unsupervised machine learning models to cope with the problem of unavailability of the annotated corpora. Finally, combinations of several sources of information have been used in current research direction.

This chapter presents brief review of the prior work in part of speech tagging however our focus has been made on part of speech taggers of Indian languages.

Automated part of speech tagging was initially explored in middle sixties and seventies [36], [49], [31]. In 1963, Klein and Simmons introduced a computational approach for grammatical coding of English words. Their primary goal was to avoid the labour of constructing a very large dictionary. Their algorithm uses a set of 30 POS categories. It first seeks each word in dictionaries, then checks for suffixes and special characters as clues. Finally, the context frame tests are applied. This algorithm correctly and unambiguously tags about 90% of the words in several pages of the Golden Book Encyclopaedia [49].

The next important tagger, *TAGGIT*, was developed by Greene and Rubin in 1971. The tag set used is very similar, but somewhat larger, at about 86 tags. The dictionary used is derived from the tagged Brown Corpus, rather than from the untagged version. This tagger correctly tags approximately 77% of the million words in the Brown Corpus [31].

First attempt of acquiring disambiguation rules from corpus were done by Hindle in 1989. In the year 1992 Eric Brill has been developed a rule based POS tagger with the accuracy rate of 95-99% [10]. POS tagging of some languages like Turkish [60], Czech [34] has been attempted using a combination of hand-crafted rules and statistical learning.

Stochastic models of part of speech tagging [25], [20], [26], [56], [55] have been widely used for simplicity and language independence of the models. Among stochastic models, bi-gram and tri-gram Hidden Markov Models (HMM) are quite popular. TNT [9] is a widely used stochastic trigram HMM tagger which uses a suffix analysis technique to estimate lexical probabilities for unknown tokens based on properties of the words in the training corpus which share the same suffix. The development of a stochastic tagger requires large amount of annotated text. Stochastic taggers with more than 95% word-level accuracy have been developed for English, German and other European languages, for which large tagged data is available. Simple HMM models do not work well when small amounts of tagged data are used to estimate the model parameters. Sometimes additional information is coded into HMM model to achieve high accuracy for POS tagging [20].

Some authors have performed comparison of tagging accuracy between linguistic and statistical taggers with favourable conclusion [16], [44].

Similarly, Indian languages like Hindi, Bengali, Punjabi, Marathi, Tamil, Telegu and Malayalam languages have many POS taggers. The oldest work on Indian language POS tagging that we found is by Bharati et al. [8]. They presented a framework for Indian languages where POS tagging is implicit and is merged with the parsing problem in their work on computational Paninian parser.

An attempt on Hindi POS disambiguation was done by Ray [65]. The part of speech tagging problem was solved as an essential requirement for local word grouping. Lexical sequence constraints were used to assign the correct POS labels for Hindi. A morphological analyzer was used to find out the possible POS of every word in a sentence. Further, the follow relation for lexical tag sequence was used to disambiguate the POS categories.

A rule based POS tagger for Tamil [4] has been developed in combination of both lexical rules and context sensitive rules. Lexical rules were used to assign tags to every word without considering the context information. Further, hand written context sensitive rules were used to assign correct POS labels for unknown words and wrongly tagged words. They used a very coarse grained tagset of only 12 tags. They reported an accuracy of 83.6% using only lexical rules and 88.6% after applying the context sensitive rules. The accuracy reported in the work, are tested on a very small reference set of 1000 words. Another hybrid POS tagger for Tamil [4] has also been developed in combination of a HMM based tagger with a rule based tagger. First a HMM based statistical tagger was used to annotate the raw sentences and it has been found some words are not tagged due to the limitation of the algorithm or the amount of training corpus. Then the untagged sentences/words are passed through the rule based system and tagged. They used the same earlier tagset with 12 tags and an annotated corpus of 30,000 words. Although the HMM tagger performs with a very low accuracy of 66% but, the hybrid system works with 97.3% accuracy. Here also the system has been tested with a small set of 5000 words and with a small tagset of 12 tags.

Shrivastav et al. [68] presented a CRF based statistical tagger for Hindi. They used 24 different features (lexical features and spelling features) to generate the model parameters. They experimented on a corpus of around 12,000 tokens and annotated with a tagset of size 23. The reported accuracy was 88.95% with a 4-fold cross validation.

Smriti et al. [70] in their work, describes a technique for morphology-based POS tagging in a limited resource scenario. The system uses a decision tree based learning algorithm (CN2). They used stemmer, morphological analyzer and a verb group analyzer to assign the morphotactic tags to all the words, which identify the Ambiguity Scheme and Unknown Words. Further, a manually annotated corpus was used to generate *If-Then* rules to assign the correct POS tags for each ambiguity scheme and unknown words. A tagset of 23 tags were used for the experiment. An accuracy of 93.5% was reported with a 4-fold cross validation on modestly-sized corpora (around 16,000 words).

Manish Shrivastava and Pushpak Bhattacharyya proposed POS tagger for Hindi based on HMM in the year 2008 [69]. Adopting rule based approach a

POS tagger for Marathi has been developed in 2006 using a technique called SRR (suffix replacement rule) by Sachin Burange et al. [13]. A Punjabi POS tagger is also developed by Singh Mandeep, Lehal Gurpeet and Sharma Shiv in 2008 with accuracy performance of 88.86% excluding unknown words [50].

As per the literature, there is a few works related to POS tagging in Manipuri and other Tibeto-Burman languages in the Indian sub-continent. In the year 2004, Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, P.K. Das have designed and implemented a morphological analyzer for Manipuri language [18]. Besides, D. S. Thoudam et al. developed morphology driven Manipuri POS tagger in 2008 [71].