Chapter I

INTRODUCTION

1.0 Introduction

Part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up words in a text corresponding to a particular part-of-speech. This marking is based on both definition, as well as context i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph. Corpus based natural language processing (NLP) tasks for popular languages like English, French etc. have been much worked on with success. On the contrary, verylittle or rather no workhas been done on languages like Kashmiri which are at the primary level in the NLP realm. One of the main reasons is the absence of annotated corpus for such languages. Corpus annotation is the practice of adding interpretative especially linguistic information to a text corpus by coding, added to the electronic representation of the text itself.

Once performed manually, methodology of POS tagging in the context of computational linguistics now involves the use of algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Part-of-speech tagging is more than just having a list of words and their parts of speech because some words can represent more than one part-of-speech at different times. For example, even ‘dogs’, which is usually thought of as just a plural Noun, can also be a Verb:

‘The sailor blows dogs’.

On the other hand, ‘dogged’, can either be an Adjective or a past-tense Verb. Just which part-of-speech a word can represent, varies greatly.

In grammar, a lexical category (also called word class, lexical class, or as part-of-speech) is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item in question. Common linguistic categories include Nouns and Verbs, among others.

Different languages may have different lexical categories, or they might associate different properties to the same one. For example, Japanese has as many as three classes of Adjectives whereas English has one, Chinese, Korean and Japanese have measure words[1] whereas European languages have nothing resembling them. Many linguists argue that the formal distinctions between parts of speech must be made within the framework of a specific language or language family, and should not be carried over to other languages or language families.

Basic grammar commonly teaches that there are 8 parts of speech in English: Nouns, Verbs, Adjectives, Prepositions, Pronouns, Adverbs, Conjunctions, and Interjections. However, there are clearly many more categories and sub-categories. For Nouns, plural, possessive and singular forms can be distinguished. In many languages, words are also marked for their ‘case’

[1] Measure words, known more formally as numeral classifiers and also called counters, count words, counter words, or counting words, are words (or morphemes) that are used in combination with a numeral to indicate the count of Nouns.

(role as subject, object, etc.), grammatical gender, and so on; whereas in most of the languages Verbs are marked for tense, aspect and other things.

The significance of part-of-speech for language processing is that it gives a significant amount of information about the word and its neighbors. This is clearly true for major categories, (Verbs Vs Nouns), but is also true for the many finer distinctions. For example, these tagsets distinguish between Possessive Pronouns ('my', 'your', 'his', 'her', 'its') and Personal Pronouns ('I', 'you', 'he', 'me'). Knowing whether a word is a Possessive Pronoun or a Personal Pronoun can tell us what words are likely to occur in its vicinity (Possessive Pronouns are likely to be followed by a Noun, Personal Pronouns by a Verb). This can be useful in most language models. A word's part-of-speech can also tell us something about how the word is pronounced. For example, the word 'content' can be a Noun or an Adjective. They are pronounced differently (the Noun is pronounced CONtent and the Adjective conTENT with the capitals indicating stress). Thus, knowing the part-of-speech is one of the basic processes of most computational linguistics related research.

Part-of-speech tagging can also be used in stemming[2] for information retrieval (IR)[3], since knowing a word's part-of-speech can help to tell us which morphological affixes it can take. They can also help an IR application by

[2]Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form generally a written word form.

[3]Information Retrieval (IR) is essentially a matter of deciding which documents in a collection should be retrieved to satisfy a user's need for information.

helping select out Nouns or other important words from a document. Automatic part-of-speech taggers can help in building automatic word-sense disambiguating algorithms, and are also used in advanced ASR[4] (Automatic Speech Recognition) language models such as class-based N-grams[5]. Part-of-speech are very often used for "partial parsing" texts, for example, for quickly finding names or other phrases for information extraction applications. Finally, corpora that have been marked for part-of-speech are very useful for linguistic research, for example to help find instances or frequencies of particular constructions in large corpora.

POS Tags carry some combination of morphological and syntactic pieces of information that is why they are also called morphosyntactic tags. In highly inflected languages, such as Greek, Latin, old English or for that matter Kashmiri, the inspection of a word out of context will reveal much about its grammatical properties. English has shed most of its inflectional features over the centuries, and the individual word will contain ambiguities that only context can resolve. Thus the '–ed' form of a Verb may be the past tense or the past participle. For some common Verbs ('put', 'shut', 'cut') the only distinction

[4] Speech recognition (Automatic Speech Recognition or Computer Speech Recognition) converts spoken words to machine-readable input. The term "voice recognition" is sometimes used to refer to speech recognition where the recognition system is trained to a particular speaker - as is the case for most desktop recognition software, hence there is an aspect of speaker recognition, which attempts to identify the person speaking, to better recognize what is being said.

[5] An N-Gram language model is a representation of an N^{th} order Markov language model in which the probability of occurrence of a symbol is conditioned upon the prior occurrence of N-1 other symbols. N-Grams are typically constructed from statistics obtained from a large corpus of text using the co-occurrences of words in the corpus to determine word sequence probabilities.

between past and present is morphologically unmarked. In many cases the distinction between Verb and Noun (love) is not morphologically marked. In order to resolve such ambiguities and make such clear distinctions different POS tagsets have been developed for English so far.

1.1 Tagset

Tagset development forms a foundation of any computational processing endeavor. The first pre-requisite for automated POS tagging is a tagset that is a set of exhaustive categories into which any token of the language can be placed. While the nature of the language is that there will always be words that are hard to classify, or are ambiguous between two categories, the tagset categories should be designed in such a way so as to minimize such problems. The fundamental problems in POS tagging task stem from the fact that a word can take different lexical categories depending on its context. The tagger has to resolve this ambiguity and determine the best sequence for a sentence.Tags are also applied to punctuation markers, thus tagging for natural language is the same process as tokenization for computer languages, although tags for natural languages are much more ambiguous.

For Indian languages, several tagsets have been developed. The most prominent among those is that developed under ILMT (Indian Language Machine Translation) guidelines, which is designed for specific languages in a flat structure capturing only coarse-level categories. Another tagset which is designed for Indian languages is that of IL-POSTS (Indian Language Part of

Speech Tagsets) hierarchical framework. IL-POSTS is a framework for Indian languages that allows language specific tagsets to be derived from it. An important consideration for its hierarchical structure and decomposable tags is that it should allow users to specify the morphosyntactic information applicable at the desired granularity according to the specific language and task.

Part of speech tagging has been studied extensively in the past two decades and lot of work has been done in various European languages including many Indian languages like Hindi, Urdu, Sanskrit, Tamil and Kannada. Comparatively speaking, little work has been done in Kashmiri in this field. The only work done in Kashmiri so far in this domain is the development of a flat tagset following ILMT (Indian Language Machine Translation) guidelines under the project "Development of Language Tools and Linguistic Resources for Kashmiri" at the Department of Linguistics, University of Kashmir.

1.2 Types of Tagsets

Tagsets can be broadly divided into two types viz:

1.2.1 Flat Tagsets

Flat tagsets cover only coarse level categories of a particular language applicable at the word level. There is no provision for modularity or feature reusability in flat tagsets. Independent labels are used for each category. Flat tagsets may be easier to process as they don't contain the higher level of granularity (list of all sub-features or attributes of the main category) and can be easily built but it is opened that such tagsets are not suitable for

morphologically rich languages. Using a Flat tagset, words are generally tagged as follows:

John\NP\ **ate**\VM\ **food**\NC\

Figure given below depicts the tag structure of above example.

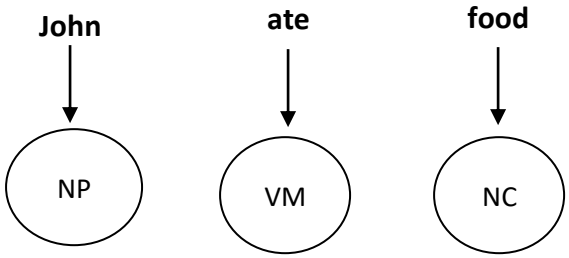

Fig. 1.1: Tag structure of Flat Tagset.

1.2.2 Hierarchical Tagsets

The categories in a hierarchical tagset are structured relative to one another (Hardie 2004). This implies that instead of having a large number of independent categories, a hierarchical tagset contains a small number of categories at the top level, each of which has a number of sub-categories in a tree structure. The morphosyntactic details are encoded in the separate layers of hierarchy, beginning from the major categories at the top and gradually progressing down to cover morphosyntactic features. This hierarchical arrangement allows the selective inclusion as well as removal of features for a specific language.

Decomposability is another desirable feature of a hierarchical tagset design as it allows different features to be encoded in a tag by separate sub-strings. A tag is considered decomposable if the string representing the tag

contains one or more shorter sub-strings that are meaningful out of the context of the original tag. Decomposable tags are believed to help in better corpus analysis (Leech 1997:27-28) by allowing to search with an underspecified search string. Using a hierarchical tagset words are generally tagged as follows:

John\N,NP,mas,sng\ **ate**\V,VM,pst,prf,fn\ **food**\N,NP,mas,sng\

Figure given below depicts the tree structure of the above example.

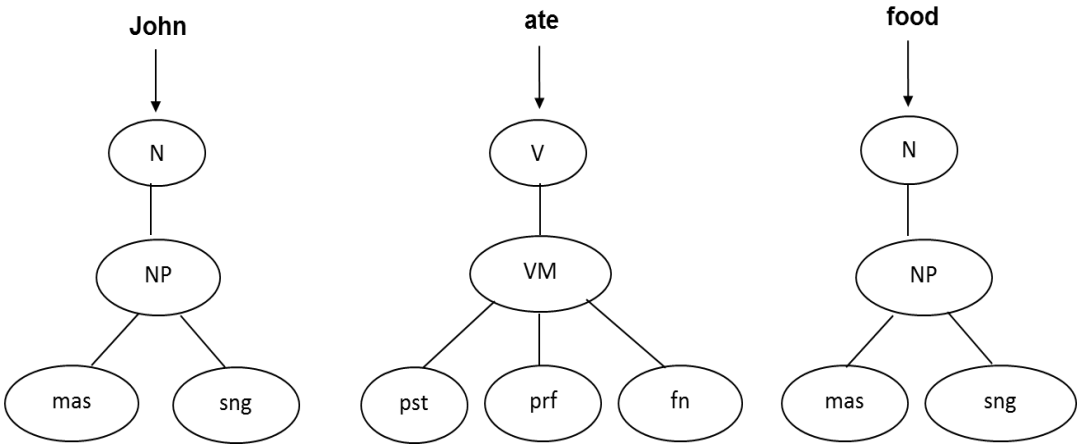

Fig. 1.2: Tag structure of Hierarchal Tagset.

1.3 Part-of-Speech Tagger

Corpus annotation is the practice of adding interpretative especially linguistic information to a text corpus by coding, added to the electronic representation of the text itself.

A typical case of corpus annotation is that of *morphosyntactic annotation* (also called *grammatical tagging)*, whereby a label or tag is associated with each word token in the text to indicate its grammatical classification. An annotated corpus serves as an important tool for investigators of natural language

processing, speech recognition and other related areas. It is a basic building block for constructing statistical models for automatic processing of natural languages. Keeping in view the importance of NLP tasks, and in order to overcome the shortage of the annotated corpus for Kashmiri an attempt is made to build an annotated corpus for Kashmiri so that the ultimate goal of developing an automatic tagger is fulfilled.

The tagger usually annotates the given word or a token. In other words, a POS tagger assigns a (unique or ambiguous) part-of-speech tag to each token in the input and then passes it to the next processing level (chunking, parsing etc.). Part-of-speech tagging is also important for corpus annotation projects, with the help of which valuable linguistic resources are created by a combination of automatic processing and human correction.

For both these applications, a tagger with the highest possible accuracy is required. The debate over the issue of which tagger solves the parts of speech problem in the best way is not over. Several approaches have been used to construct automatic taggers. Most of the work done is based on statistical methods using n-gram models for Hidden Markov Model–based tagger (Church 1988; De rose 1988; Cutting et al 1992; Merialdo 1994; Kupiec 1992; Brill 1992 and Voutilainen et al 1992). In these approaches a tag sequence is chosen for a sentence that maximizes the product of lexical and contextual probabilities as estimated from a tagged corpus.

Broadly speaking, taggers are divided into two categories.

1.3.1 Rule Based Tagger

Rule based taggers generally involve a large database of handwritten disambiguation rules which specify, for example, that an ambiguous word is a Noun rather than a Verb if it follows a Determiner. The earliest algorithms for automatically assigning part-of speech were based on two stage architecture (Harris 1962; Klein and Simmons 1963; Green and Rubin 1971). The first stage uses a dictionary to assign each word a list of potential parts-of-speech. The second stage uses large lists of handwritten disambiguation rules to winnow down this list to single parts-of-speech for each word (Jurafsky and Martin 2002).

The ENGTWOL tagger (Voutilainen 1995) is based on the same two stage architecture, although both the lexicon and the disambiguation rules are much more sophisticated than the early algorithms. The ENGTWOL lexicon is based on the two-level morphology and has about 56,000 entries for English word stems (Heikkila 1995), counting a word with multiple parts-of-speech (e.g. nominal and verbal sense of 'hit') as separate entries, and not counting inflected and many derived forms. Each entry is annotated with a set of morphological and syntactic features.

1.3.2 Stochastic Part-of-Speech Tagger

Stochastic taggers use probabilities in tagging. Probabilities in tagging were first used by Stolz et al 1965. A complete probabilistic tagger with Viterbi decoding was sketched by (Bahl and Mercer 1976), and so on. Various

stochastic taggers were built in the 1980s (Marshall 1983; Garside 1987; Church 1988; DeRose 1988). Stochastic tagging generally uses a tagging algorithm known as Hidden Markov Model or HMM tagger. All stochastic taggers work on the basis of one simple generalization of "pick the most-likely tag for this word" approach (Jurafsky and Martin 2002).

One of the popular stochastic POS taggers is TnT tagger which is observed to show high accuracy in English and some other languages. TnT, the short form of *Trigrams'n'Tags,* is a very efficient statistical part-of-speech tagger which is trainable on different languages and virtually on any tagset. The component for parameter generation trains on tagged corpora. The system incorporates several methods of smoothing and handling unknown words. TnT is not optimized for any particular language. Instead, it is optimized for training on a large variety of corpora and it is very easy to adapt the tagger to a new language, new domain or new tagset and these positive features strongly favored the use of TnT for the present work. Moreover, TnT is optimized for speed. For part-of-speech tagging TnT uses second order Markov Model. An important characteristic of TnT tagger is that it not only assign tags to words but also to the probabilities. Average accuracy of the tagger is 95% - 97%, depending on the language and the tagset.

1.3.3 Transformation Based Tagger

Transformation based tagging, sometimes called Brill tagging, is an instance of the Transformation Based Learning(TBL) approach to machine

learning(Brill,1985), and draws inspiration from both the rule-based and stochastic taggers. Like the rule based taggers, TBL is based on rules that specify what tags should be assigned to what words. But like the stochastic taggers, the TBL is a machine learning technique, in which rules are automatically induced from the data. Like some but not all of the HMM taggers, TBL is a supervised learning technique; it assumes a pre-tagged training corpus.

1.4 Motivation

Looking at languages from a computational perspective, developing a Part of Speech tagger is one of the primary pursuits of Natural Language Processing (NLP). This becomes more important in the case of Kashmiri where computational work has only recently begun. Developing a Part of Speech tagger for Kashmiri with optimum level of accuracy would be a significant contribution because it would lead to its use in applications like Machine Translation, Information Extraction, Information Retrieval, Lexicography, Spelling and Grammar Checker, Morphological Analyzer, etc.

POS tagging can be both manual as well as automatic. Manual tagging, though more accurate, is a time-consuming, long and continuous process. Hence, the automatic tagger is essential to speed up the process of POS tagging with less chance of errors and inconsistencies. Various automatic POS taggers have been developed worldwide using linguistic rules, stochastic models and hybrid taggers (a combination of both). Different kinds of taggers have certain advantages as well as disadvantages. Automatic tagging is a challenge for Indian languages which are highly inflectional and morphologically rich. Hence, the development of high accuracy POS taggers is a challenging task.

1.5 Objectives

The primary objective of the thesis can be summarized as follows:

1. To study the different approaches to part of speech tagging.
2. To study computational linguistics approaches and their applications in part of speech tagging.
3. To study and analysis the linguistics features of the proposed language.
4. To study different rules/guidelines for developing tagset.
5. To develop Part of Speech Tagset for Kashmiri.
6. To develop an automatic Part of Speech tagger for Kashmiri.

1.6 Methodology

1. A general overview of Part of Speech Tagging approaches was first obtained.
2. Study of some computational techniques of Part of Speech Tagging and their applications in the field of natural language processing.
3. Study of linguistics features of Kashmiri and collection of text data (Corpus).
4. A hierarchical tagset for Kashmiri was developed by following the Eagles Guidelines and the Penn tree bank tagset. Many other Indian tagging guidelines like IL-POST, ILMT and Sanskrit tagset were also taken into consideration.
5. An Automatic Part of Speech tagger for Kashmiri was developed to generate the tagged output with high accuracy level.

1.7 Main Contributions

1. Studies and analysis of various POS tagging algorithms and computational linguistics approaches.
2. Studies and analysis of Linguistic features of Kashmiri.
3. Development of a tagset for Kashmiri taking into consideration both language features in general and the idiosyncratic features.
4. Development of annotated corpus.
5. Development of POS tagger for Kashmiri.

1.8 Thesis Outline

Chapter II:

This chapter presents an overview of Part of Speech Tagging, its different paradigms and standard approaches. It also describes important guidelines and framework for developing Part of Speech Tagset.

Developing a Tagset is a prerequisite for natural language processing of any language. Thus we can say prior to the development of an automatic tagger it is necessary to build suitable tagging guidelines for any language, so that tagged corpus for any language can be built.

Chapter III:

This chapter describes the framework for developing Kashmiri Part of Speech Tagset. For designing a Kashmiri tagset, apart from following the Eagles Guidelines and the Penn tree bank tagset, many other Indian tagging guidelines

like IL-POST, ILMT and Sanskrit tagset were taken into consideration. The tagging schema for Kashmiri is designed taking into consideration both language features in general and the idiosyncratic features of Kashmiri. After careful consideration a hierarchical tagset was favored. The whole design of the tagset developed so far revolves around three distinct features into which the grammatical schema is distributed. The features are:

I. Category
II. Type
III. Attribute

Categories involve major grammatical categories like Nouns, Verbs etc. The type includes the type of those grammatical categories like Common Noun and Proper Noun for Noun category, Main Verb, Auxiliary Verb etc. for Verb category, and so on. The attribute level takes features within each type like Gender (masculine, feminine), Number (singular, plural), Case (dative, ergative, ablative, etc), Tense and Aspect etc into consideration. The category list includes all Kashmiri categories that can occur. The type list within a category includes all types of the category that can occur. The attribute list includes all possible attributes of the type that can occur.

The overall number of Category Tags used in the proposed tagset is 26, Type tags are 21 with their corresponding attributes.

Chapter IV:

This chapter begins with the description of Hidden Morkov Model (HMM) and we have used HMM for automatic POS tagging of natural language text. The HMM models use the following three sources of information.

a. *Symbol emission probabilities,* i.e. the probability of a particular tag t_i, given a particular word w_i, $P(w_i \mid t_i)$.
b. *State transition probabilities,* i.e. the probability of a particular tag depending on the previous tags, $P(t_i \mid t_{i-1}t_{i-2}.....t_{i-k})$.
c. *Probability for the initial sate,* i.e. the probability of a particular tag as an initial state of a Markov model

We have implemented Hidden Morkov Model to understand the complexity of the POS tagging task. In this model the tag probabilities depend only on the current word: The effect of this is that the each word in the test data will be assigned the tag which occurred most frequently for that word in the training data.

The experiments were conducted with five different sizes (100K, 120K, 140K, 160K and 180K words) of the training data to understand the relative performance of the models as we keep on increasing the size of the annotated data.

Initially corpus of around 2,00,000 words was taken. Then this corpus was divided into two parts out of which 50% was used for training the tagger and the remaining 50% of the corpus was used as test data for checking the accuracy of the tagger. The test data is unseen during training. By using these proportions the overall accuracy of 62.55% was found. Then separate accuracies of known and unknown words were also calculated. In the test the accuracy of known and unknown words was 87.02% and 38.19% respectively.

Each result was obtained by repeating the experiment 5 times with different partitions that is, the first partition was taken as 50%- 50% then next

partition was 60% - 40% and so on and so forth to check the accuracy of the tagger. The results obtained are shown in the table given below:

Corpus			Overall		Known Words		Unknown Words	
Total Corpus	Training Data	Test Data	Acc. (%)	Errors (%)	Acc. (%)	Errors (%)	Acc. (%)	Errors (%)
2,00,000	**1,00,000**	**1,00,000**	**62.55**	**37.45**	**87.02**	**12.98**	**38.19**	**61.81**
	1,20,0000	**80,000**	**68.10**	**31.90**	**85.22**	**14.78**	**40.15**	**59.85**
	1,40,000	**60,000**	**76.39**	**23.61**	**84.78**	**15.22**	**42.11**	**57.89**
	1,60,000	**40,000**	**85.64**	**14.36**	**85.62**	**14.38**	**51.68**	**48.32**
	1,80,000	**20,000**	**96.28**	**03.72**	**87.62**	**12.38**	**52.35**	**47.65**

Table1.1: Accuracy of Kashmiri POS tagger.

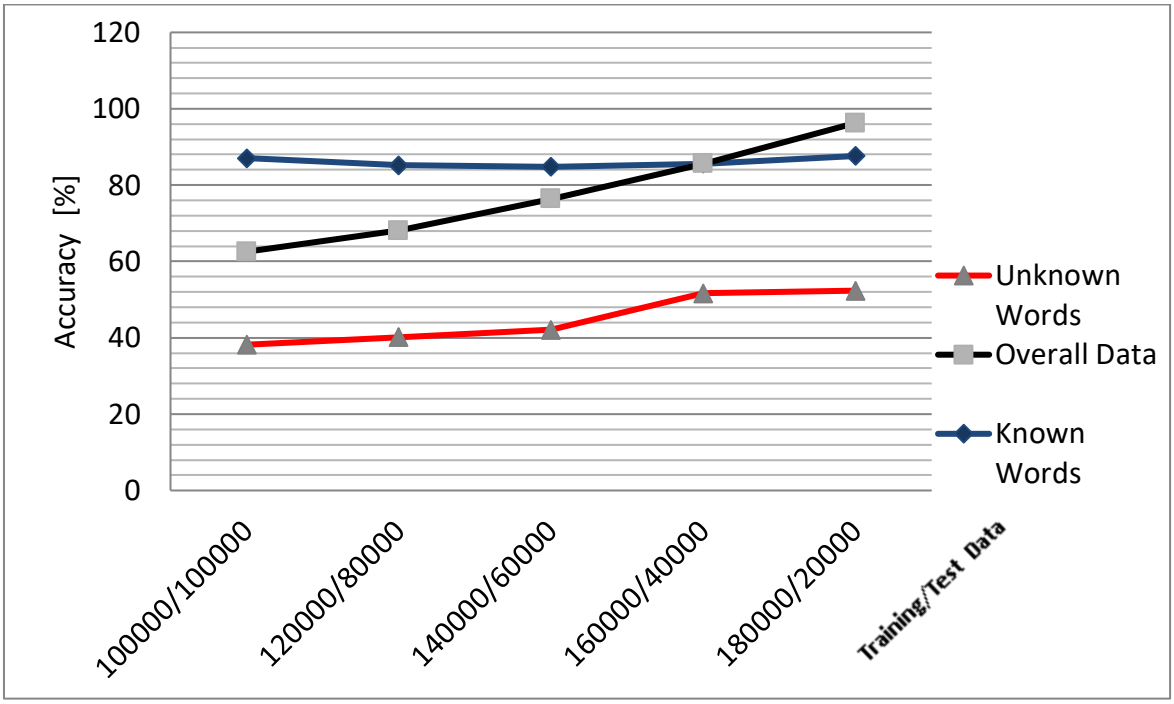

Fig 1.3: Kashmiri Corpus: POS learning curve.

Chapter V:

Finally, this chapter presents the conclusion. Summary of the works and contributions are outlined along with a discussion on scope for future research work.

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. Developing a POS tagger is the first attempt towards building a natural language processing (NLP) tool for Kashmiri.. Accuracy of the present tagger is about 96.28% which would be increased by retraining the tagger by using more and more corrected data.

1.9 Conclusion

POS tagging is typically achieved by rule-based systems, probabilistic data-driven systems, neural network systems or hybrid systems. For languages like English or French, hybrid taggers have been able to achieve success percentages above 98% (Schulze et al 1994). In part-of-speech tagging by

computer, it is typical to distinguish around 50 to 150 separate parts of speech for English, for example, NN for singular Common Nouns, NNP for plural Common Nouns, NP for singular Proper Nouns and so on.