Chapter 1

Introduction

This thesis deals with a common problem in the area of Natural Language Processing: How can we automatically assign parts-of-speech to words in a text? Assigning a POS tag to each word of an unannotated text by hand is very time consuming, which results in the existence of various approaches to automate the job. Automated POS tagging is a technique to assign with its appropriate lexical categories [7]. The process takes a word or a sentence as input, assigns a POS tag to the word or to each word in the sentence and produces the tagged text as output. Part of speech tagging is a preliminary and important component of computational linguistics or natural language processing [3]. Human languages are generally known as natural languages; the science of studying natural languages falls under the area of linguistics [4] and its implementation using computational means is regarded as computational linguistics. It has been predicted that, in the future, the main method of communication between humans and computers (or other processing devices) will be natural language (NL), in both spoken and written forms. In fact, today, humans already use NL for interacting with various systems, i.e. writing queries to search engines and entering credit-card numbers using voice. Furthermore, various systems have already been developed to process NLs for some particular task, e.g. grammar correction, information extraction, corpus annotation and machine translation. Part-of-speech tagging was also considered an integrated part of natural language processing [8], because there are some cases where the accurate part of speech cannot be determined if the semantics or even the pragmatics of the context is not understood.

Research on part-of-speech tagging has been inherently linked to corpus linguistics. The first major corpus of English for computer analysis was the Brown Corpus developed at Brown University by Henry Kucera and W. Nelson Francis, in the mid-1960s[9][10].Initial attempts at Part of tagging was done [9][10] in two stages (Harris, 1962, Klein and Simmons, 1963; Rubin and Greene 1971). The tagger TAGGIT (Greene and Rubin, 1971) was used for initial tagging of the Brown Corpus [9][10].The ENGTWOL Tagger which was developed in year 1995 is based on constraint grammar architecture [10], and this tagger can be considered most important in the direction of this field. These taggers generally use rule-based models where the rules are manually written by linguistics experts. The advantage of rule model is that the rules are written from a linguistic point of view and the rules can be used to collect complex kinds of information. Due to this reason, it allows constructing the tagger with an extremely high accuracy. However, handling all rules for tagger is not an easy task and thus demands expertise. Thus the context frame rules have to be developed with the help of language experts. Further, if anyone is to transfer or translate the rule-based tagger to another language, it would involve starting from scratch again.

A comprehensive set of linguistically motivated rules or a large annotated corpus is a precondition for the development of an automatic POS tagger. However, development of such corpora and rules has been very limited to only a few languages like English and few other languages. But POS taggers for Indian languages are not readily available due to lack of such rules and also because of inadequate amount of annotated text. Manual part of speech tagging is a tedious and difficult process. As such, certain methods have been devised so that with small amount of tagged resources, an effective part of speech tagging task can be carried out.

The proposed methodology is applicable for the POS disambiguation task of any poor language. We have looked at working with certain standard learning approaches that can function with only a scarce data. So, we have also carried out comparative studies of different part of speech (POS) tagging techniques and their accuracies. We also studied and analyzed the effect on the learning algorithms of POS tagging using different features.

1.1 Parts of Speech Tagging

Part-of-speech tagging is the process of classifying or labeling the words in a text with their appropriate part of speech. Part of speech tagging [40], is also called word-category, disambiguation or grammatical tagging. POS tagging is also based on both the definition and its context i.e. the relationship of the word with its adjacent words and related words in the phrase, sentence or paragraph that it exists in. A simplified version, Part of Speech Tagging (POST) is defined as the identification of words as nouns, pronouns, adjectives verbs, adjectives, etc. in a sentence. Part Of Speech (POS) tagging can be considered as a simplified form of morphological analysis and morphological analysis deals with finding the internal structure of the word in a sentence; whereas POS Tagging deals only with assigning or labeling an appropriate POS tag to the word.

Part-of-speech tagging is not only classified or identified list of words and their parts-of-speech; it also can represent some words as more than one part-of-speech at different times. For example, even ‘Plants’, which is usually thought of as just a plural Noun, can also be a Verb:

‘Plants/N need light and water.

"Each one plants/V one sapling."

Part of Speech(POS) tagging is the first step towards the machine translation or natural language understanding, it is very important to achieve a high level of accuracy, which otherwise may alienate further stages of the natural language understanding. Part-of–speech tagging is also applied for a number of applications, including – speech synthesis, recognition, information extraction, partial, machine translation, and lexicography, etc. In the following, the application of POS tagging has been briefly discussed.

Speech synthesis and recognition: POS tagging can be used for collecting information about the particular word and their closest counterparts which can be useful for speech recognition in a language model. Part-of-Speech (POS) tagging is also useful for the pronunciation depending on the grammatical category.

***Information retrieval and extraction*:** Part-of-speech is also used for more refined and accurate information extraction and retrieval.

Machine translation*:* The probability of accurately translating a word in the source language into a word in the target language is positively dependent on the Part-of-speech (POS) category of the source word.

Higher-level syntactic processing: Tagging often serves as a precursor to higher-level syntactic-processing systems. e.g. noun phrase chunkers (programs to find NPs in each sentence) use a combination of word and POS information to learn either regular expressions for NPs in a sentence that are likely to indicate the beginning or ending of a phrase. Tagging can also be important for speech synthesis. e.g. the word record is pronounced differently depending on whether it is a noun or a verb.

1.2 Problems of POS Tagging

There are two major problems of POS tagging: ambiguous words and unknown words. The first problem is the existence of words for which more than one tag is possible. The solution of the problem is taking the context into consideration rather than single words. This is a trivial task for humans but not so easy for automatic text taggers. Below is an example sentence, which has different possible tags for one word:

We can can the can. ……..(1.1)

When will the race end?.....(1.2)

In the above example (1.1), 'can' corresponds to auxiliary verb, verb and noun respectively. Part-of-speech tagging is the process of assigning to each word in an input text a proper syntactic tag or part of speech tag in its context of appearance. In most cases, the ambiguous word can be disambiguating completely using the adequate context as in above example (1.2). The word, 'race' would be disambiguated as noun because its previous word 'the' is determiner; ambiguity is resolved by simply looking at the previous tag. But it is not sufficient to disambiguate the word by such simple context and may require much more language knowledge. The most challenging problem in POS tagging disambiguation is to determine the proper context and adequate features.

The latter problem is the occurrence of unknown words, i.e. words that do not exist in the corpus. Thus, it is a significant design issue to provide mechanisms for handling unknown words.

1.3 Motivation

Part-of-speech (POS) tagging is very important for various text understanding applications including machine translation. Machine translation (MT) is a sub-field of computational linguistics which is used to translate text or speech from one language to another. But MT alone usually cannot give a good result of a text translation because recognition of words, phrases and their neighbour in the target language is not an easy task. This problem can be solved to some extent with corpus, statistical techniques and Part of speech tagging and these are the rapidly growing fields for better translations.

Modern NLP technology can be used for preserving language and other computational work. This becomes more important in the case of Nepali where computational work has only recently begun. Developing a Part of Speech tagger for Nepali 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.

Part-of-speech (POS) tagging can be solved in two ways: manual as well as automatic. In Manual tagging the accuracy of POS tagging is more, but it takes lot of time and is a continuous process. Hence, the automatic tagger is the preferred choice to speed up the process of Part-of-speech tagging with less chance of mistakes and inconsistencies. There are many automatic POS taggers which have been developed worldwide using linguistic rules, stochastic models and hybrid model (combination of both). Different types of taggers have their own advantages as well as disadvantages. Hence constructing

Automatic part of speech tagging is a challenging task especially for Indian languages, which are highly morphologically rich and inflectional.

Therefore, those are the fundamental requirements to develop an automatic part of speech tagger in Nepali for the overall development of the language in the field of Computational Linguistics. With this motivation, we identify objectives of this thesis.

1.4 Objective

The main objective of this dissertation is to build a Nepali POS tagger based on Hybrid framework. The other objectives of the thesis can be summarized as follows:

a) To study the paradigms and different approaches to part of speech tagging.
b) To study some computational linguistics approaches and their applications in part of speech tagging.
c) To study and analyse the linguistic features of the proposed language.
d) To study different rules/guidelines for developing tagset.
e) To develop Part of Speech Tagset for Nepali.
f) Manual part of speech tagging is a tedious and difficult process. As such, certain methods have been devised so that with small amount of tagged resources, an effective part of speech tagging task can be carried out.

1.5 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 Nepali and collection of text data (Corpus).
4) A hierarchical tagset for Nepali was developed by BIS Standard and the Penn tree bank tagset guidelines. 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 Nepali was developed to generate the tagged output with high accuracy level.

1.6 Thesis Organization

Chapter 2: This chapter presents brief review of the prior work in part of speech tagging and part of speech taggers in Indian languages as well as International status on POS tagging. Due to the huge number of publications in this field, it is very difficult to give a comprehensive review of the Part of speech tagging and its related work. Another difficulty of review is the diverse language dependent works based on several theories, methodologies and techniques used by researchers over the years. We have briefly reviewed of different part of speech (POS) tagging techniques/ methodologies and their accuracies. We also focus onto the detailed review of the Indian language POS taggers.

This chapter presents brief review of the prior work in part of speech tagging and part of speech taggers in Nepali languages.

Chapter 3: This chapter presents an overview of different Techniques or paradigms for part-of-speech tagging. We also describe our approach to eliminate part-of-speech ambiguity It also presents some applications of part of speech tagging in the field of computational linguistics. Finding Unknown words is a difficult task and the problem of unknown words is also discussed in this chapter.

Chapter 4: This chapter gives information about several important issues related to Corpora Collection. This chapter presents brief review of Nepali Tagset and Description of Nepali Tagset.

Chapter 5: This chapter describes the proposed architecture on Nepali POS tagging. In this chapter, we also present the Proposed Algorithm and Graphical User Interface Tool.

Chapter 6: This chapter presents experimental results and evaluation on Nepali POS tagging using Hybrid approach. We also present the uses of Input/output of the Program and their effective performance or accuracy.

Chapter 7: Finally, this chapter presents the conclusion, summary of the work and contributions along with a discussion on scope for future research work.