|  |
| --- |
| **Chapter 1: Introduction** |

Part-of-speech (POS) tagging, also known as grammatical tagging and it is the common form of corpus annotation. POS tagging is the one of the main components of Natural Language Processing (NLP). POS tagging label the word with their appropriate Parts-of-Speech and then it expresses the formation of words in a sentence. In computational application POS tags usually help us for develop different NLP techniques (e.g. sentimental analysis, speech recognition) and it also maintain a link between human and modern artificial machines.

* 1. **Objective**

This paper developed a POS tagging technique for online Bengali newspaper-based data source. A supervised rule-based technique was proposed for identify Bangla sentence and its corresponding POS tags. Our main goal is to tag the parts of speech of a Bangla word which based on the pre-annotated corpus and train the dataset for extend its functionalities. Our paper was mainly focused on the sports-based news corpus. We have also identified the characteristics of Bengali grammar for this domain and several limitation and work procedure will be discussed later in this paper.

* 1. **Background of Study**

Bengali also known by its endonym Bangla (বাংলা), is an Indo-Aryan language primarily spoken by the Bengalis in the Indian subcontinent and the whole part of the Bangladesh [24]. Geographically Bengali language is one of the most popular language in the world. Around approximately 250–300 million people in the world speaks with this language [25]. Except some part in the Indian continent more than 90% people of Bangladesh are speaks with Bangla language. Although use of vast majority of this language is still popular and ranked 7th in the languages of native speakers but still we are far behind from the modern science of language development process.

Part-of-speech categorization is taught in school-age children to construct a meaningful sentence. Bangladesh is a sports-oriented country where 80% of the people’s daily conversation contains related to sports. So, in this case we have narrowed down our working domain in sports section of Bengali newspaper. Wh

* 1. **Scope**

Our thesis scope was bound to find a POS tagging mechanism from which we can find out the POS tags against e sentence. We focused to determine that the domain we choose whether it was correct for this particular area or not. Traversing the paper, we will be discussing how Bengali sentence for a news corpus can be annotated and the different findings that we found analyzing corpus. We will be trying to demonstrate the behavior of a Bengali sentence in the sports domain, how they occur, how they are placed and vary from each other and list out the findings on that and come to a decision against the findings we got.

* 1. **Brief summary of Parts-of-Speech tagging**

Part-of-Speech explains how a word is used in a sentence. There are several grammatical categories in a POS tagger. A POS tag is a special label assigned to each word in a text corpus. POS tags are used in corpus searches and in text analysis tools and algorithms. A corpus is a large collection of texts which is usually labelled with information about part-of-speech and grammatical category. POS tagging is used in natural language processing (NLP) which is mainly use for machine learning. POS tags are extremely useful since they provide linguistic signal on how a word is being used within the scope of a phrase, sentence, or document.

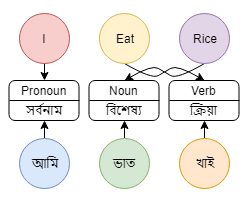


Figure 1-A: POS Tagging

A POS consists different types of grammatical categories such as tense, number etc. All kind of POS tags have a several different tag set and different language tag sets are typically different from each other. Grammatical categories are subdomain to POS tag set. Tags are useful for building parse trees which shows the relations between words. Some set of specific rules, conventions, and principles which dictate how phase, clause and words combined into sentences. There are some lexical categories which words are assigned based on their syntactic context and role. The most common part-of-speech are: noun (বিশেষ্য), pronoun (সর্বনাম), adjective (বিশেষণ), verb (ক্রিয়া), adverb (ক্রিয়া বিশেষন), preposition (অব্যয়).

There are several POS tags set available to build a linguistic corpus. Penn treebank tag sets are very much popular and it is widely used for building POS tagger. Besides that, it got some sub categories too. In part-of-speech tagging by computer, it is typical to distinguish from those sub categories. For simplicity we will restrict to the common part-of-speech that are known to general people. Identifying part of speech tags is much more complicated than simply mapping words to their part of speech tags. This is because POS tagging is not something that is generic. It is quite possible for a single word to have a different part of speech tag in different sentences based on different contexts. That is why it is impossible to have a generic mapping for POS tags.

Sometimes it is not the same case that one word represents one part-of-speech. Multiple words can represent a single POS tag and each of them also represent their own POST tag as individually they are different but with combined format it may indicate different meaning. It is a common finding that the ‘Verb’ tag almost always consists of several words. As is represent the act of a subject, it can’t always be representable by a single word. And therefore, multiple words need to consist a verb.

POS-tagging algorithms fall into two distinctive groups:

* **Rule-Based POS Taggers**
* **Stochastic POS Taggers**

**Rule-Based Tagging**

Automatic part of speech tagging is an area of natural language processing where statistical techniques have been more successful than rule-based methods.

Typical rule-based approaches use contextual information to assign tags to unknown or ambiguous words. Disambiguation is done by analyzing the linguistic features of the word, its preceding word, its following word, and other aspects.

**Stochastic Part-of-Speech Tagging**

The term ‘stochastic tagger’ can refer to any number of different approaches to the problem of POS tagging. Any model which somehow incorporates frequency or probability may be properly labelled stochastic.

The simplest stochastic taggers disambiguate words based solely on the probability that a word occurs with a particular tag. In other words, the tag encountered most frequently in the training set with the word is the one assigned to an ambiguous instance of that word. The problem with this approach is that while it may yield a valid tag for a given word, it can also yield inadmissible sequences of tags.