|  |
| --- |
| **Chapter 2: Background Research Review** |

* 1. **Previous Research**

There are several different techniques used for POS tagging. Both supervised and unsupervised [10], [1] model of the POS tagging techniques were used in previous researches. Hammad [1] proposed an unsupervised way of building a POS tagger. The author claims that supervised model is time consuming rather than an unsupervised model and Total 54 tag set was used. The author used the pre-tagged corpus and in and it follows the Baum-Welch algorithm which a modified model of HMM but the performance evaluation was not shown in this paper. Rule based POS tagging [2], [3], [15], [18] also shows the significant improvement for Bangla NLP research. CFG and CSG grammar-based parts-of-tagging [3], [6], [9] technique also verified by some authors. We have learned that for Bangla we cannot identified any Bengali sentence just consider the simple verb and simple noun [2]. Beside identifying Bangla sentence [2] proposed and well developed a very good CFG rule to identify the syntax of Bangla sentence. CFG and CSG categories have difference which is showed in [3] and it need to be considered before start of making a rule-based parts-of-speech tagger. Debsari [4] proposed a 4 layer of hybrid parts-of-speech tagging for Bangla which will achieve accuracy closer to the English and French language and also proposed a rule-based system for detect word ambiguity for noun.

Stemming and Lemmatization are the most important part of reducing inflectional and derivationally related forms of a word. In [20], a language independent and supervised model was developed using FIRE Bengali corpus and Tagor’s short stories which identifies the lemma of Bangla sentence. Their process involves of Word sense disambiguation with the result of 69.57% accuracy. Newspaper based Bengali corpus also involved in developing for identify stem of words [5]. Total 1000 sentence along with 10 tag set they [4] achieve 74.6% accuracy. A rule based [18] stemming process which achieves 88% accuracy but is limited to verb and noun. The author implied that it can be extensible for other parts-of-speech tags.

In [22], different types of language model for Bangla parts-of-speech tagging was shown. With the 41 tag set and 4048 token the rule-based Brill model gives better performance for making parts-of-speech tagging corpus for Bangla. Although statistical based [21] method with 26 POS tags and 72341 wordform gives more accuracy (90.3%) than Brill method. Hybrid model of parts-of-speech tagging [10] is far better than any other of method we have analyzed in our literature review. Using Hidden Markov Model corpus-based approach was followed and it give 96.28% accuracy. For development of POS tagging Indian language have a specific corpus named Anncorra [16] which is gives guideline for make any rule based parts-of-tagging system in any Indian language. In [10], it also reviewed that for supervised technique we need a pre-annotated corpus which is rather optional in

unsupervised technique. But if a POS tagger make with CFG approach it may give us a good result in morphological analysis but word ambiguity probability is much higher in this process [6].

Universal Networking Language (UNL) is a declarative formal language specifically designed to represent semantic data extracted from natural language texts [26]. The UNL is a common language to exchange information through computers which can deal with natural languages [27]. In [7], a inflectional morphology was proposed to produce word from another words maintain the same POS category. Using UNL word sense ambiguity detection also possible to identify [8]. Correction of Bangla grammar can be a difficult task in machine translation. But CFG based grammar identification [15] using HPSG structure can overcome this problem to make error free grammar rule for POS tagger.

* 1. **Related Work**

Research of Bengali newspaper-based corpus [11], [12], [13] have been done previously with maximum accuracy of 91.23% accuracy [13] using SVM. In [13], 34 million wordform has been recorded with total 108305 of news document. Total 26 POS tags were used for lexicon development and both SVM and HMM method used to build this POS tagging. Another research [12] which used knowledge-based AI technique which contains 74698 words form. This research is identical to our thesis and but the word forms are low in size compared to our corpus. Both [11] and [12] used Prothom-Alo as their primary data source but in [11] author compares the web-based news corpus with the CIIL corpus and they found the less amount of foreign words in their statistical analysis. Although [11] have done only the statistical analysis and they have not proposed any model for identify the accuracy level of their corpus.