ABSTRACT

The main aim of this thesis is to build a Bangla Parts-of-Speech (POS) tagger using a Hybrid Approach and demonstrate that a Hybrid POS tagger performs better than any tagger based on a single approach like a statistical or linguistically-motivated approach.

The hybrid approach in this thesis is framed by combining statistical and linguistically-motivated tagging techniques, namely, Conditional Random Field (CRF), Rule-based and Look-Up algorithms. In order to defend the argument, it is aimed to merge the three algorithms through a corrective approach and build a distinct Hybrid tagger. The output of the CRF tagger is compared with the output of the Hybrid tagger to compute the difference in their accuracy levels to prove that the hybrid approach is more competent.

The development of the Hybrid POS tagger for Bangla includes four modules – CRF module, Rule-based tagging module, Look-up tagging module and Disambiguation module. A Bangla corpus consisting of news articles collected from the TDIL corpus and an online newspaper ‘Anandabazar Patrika’ is used to test the Hybrid POS tagger. The corpus comprises of One hundred and eighty seven files of news articles blended together into a single Unicode compliant text file.

The entire corpus is divided into two sets and pre-processed for using in the proposed tagger. A manually annotated training set consisting of 35,216 tokens is used for training data in the CRF module and a corpus of 3,76,283 tokens is tagged by the
CRF module. The output of the CRF tagger acts as the input to the Look-up tagging module. The Look-up tagging module constitutes a dictionary of closed class of unambiguous words and a program in Perl to assign tags to all the words belonging to closed class in the input by matching.

The Rule-based tagging module constitutes the third stage where disambiguation rules are crafted using the unambiguous inflectional suffixes of the word classes of the open class category, i.e., nouns, verbs, adjectives and adverbs. The consolidated output of the CRF and Look-up systems is fed as the input to this module which produces tagged output using the inflectional rules.

The final module is the disambiguation or correction module which compares the output of all the three modules and prints a single tagged output. The tags produced by the look-up tagging module are the most preferred ones followed by the tags obtained from the rule-based tagging module. In other words, the output of the look-up tagging module and rule-based tagging module are used to correct the output of the CRF tagging module.

Finally, the outputs of the CRF tagging module and Hybrid POS tagger are compared and the performance of these two taggers is compared using the evaluation application of TnT model. The Hybrid POS tagger has produced an accuracy of 86.47%, while the CRF tagger an accuracy of 75%, which proves the initial hypothesis.