World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:15, No:12, 2021

Automatic Tagging and Accuracy in Assamese Text Data

Authors : Chayanika Hazarika Bordoloi

Abstract : This paper is an attempt to work on a highly inflectional language called Assamese. This is also one of the national languages of India and very little has been achieved in terms of computational research. Building a language processing tool for a natural language is not very smooth as the standard and language representation change at various levels. This paper presents inflectional suffixes of Assamese verbs and how the statistical tools, along with linguistic features, can improve the tagging accuracy. Conditional random fields (CRF tool) was used to automatically tag and train the text data; however, accuracy was improved after linguistic featured were fed into the training data. Assamese is a highly inflectional language; hence, it is challenging to standardizing its morphology. Inflectional suffixes are used as a feature of the text data. In order to analyze the inflections of Assamese word forms, a list of suffixes is prepared. This list comprises suffixes, comprising of all possible suffixes that various categories can take is prepared. Assamese words can be classified into inflected classes (noun, pronoun, adjective and verb) and un-inflected classes (adverb and particle). The corpus used for this morphological analysis has huge tokens. The corpus is a mixed corpus and it has given satisfactory accuracy. The accuracy rate of the tagger has gradually improved with the modified training data.

Keywords : CRF, morphology, tagging, tagset
Conference Title : ICLAL 2021 : International Conference on Linguistics and Applied Linguistics
Conference Location : Paris, France
Conference Dates : December 30-31, 2021