Abstract

Part Of Speech (POS) tagging is the most important pre-processing step in almost all Natural Language Processing (NLP) applications. It is defined as the process of classifying each word in a text with its appropriate part of speech. In this paper, the probabilistic classifier technique of Maximum Entropy model is experimented for the tagging of Kannada sentences. Kannada language is agglutinative, morphologically very rich but resource poor. Hence 51267 words from EMILLE corpus were manually tagged and used as training data. The tagset included 25 tags as defined for Indian languages. The best suited feature set for the language was finalised after rigorous experiments. Data size of 2892 word forms was downloaded from Kannada websites for testing. Accuracy of 81.6% was obtained in the experiments which prove that Maximum Entropy is well suited for Kannada language.

References

A Maximum Entropy Approach to Kannada Part Of Speech Tagging


Index Terms

Computer Science
Artificial Intelligence
Keywords
Natural Language Processing  Part Of Speech Tagging  Maximum Entropy