Abstract

Language identification is an important pre-processing step for any Natural Language Processing task. Kannada Language is an Indian Language and lot of research is being carried out on Kannada Language Processing. Major parts of online documents like websites are combination of Kannada and English Sentences. Language Identification is a preprocessing step for NLP tasks like POS tagging, Sentence Boundary Detection or Data mining technique. In this paper, we present an n-gram method of language identification for documents with Kannada, Telugu and English sentences. It has been shown how performance can be improved by n-gram processing only last word of the sentence instead of complete sentence. This method could also be preprocessing step for Sentence Boundary Detection discussed in [1].

References


B. Ahmed, S. Cha, "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", Proceedings of CSIS 2004, Pace University, May 7th, 2004

http://software.wise-guys.nl/libtextcat/

Index Terms

Computer Science

Algorithms

Keywords
N-gram Processing  Verb Suffix  Language Identification.