Sentence Boundary Detection in Kannada Language

Abstract

Sentence Boundary Detection is a pre-processing step for any Natural Language Processing application. Various algorithms have been used to achieve Sentence Boundary Detection or Disambiguation in different languages. In this paper, a rule based method is proposed and tested to achieve Sentence Boundary Detection for Kannada Language. Kannada being a grammatically rich Indian language is analyzed based on semantics and tested with a 227K bytes corpus. The code is written in C using wide characters, with support for Unicode. Results showed 99.2% success in detecting sentence boundary.

References

disambiguation. Computational Linguistics 23 241–267
  - Agarwal N., Ford K., and Shneider M., Sentence Boundary Detection using a MaxEnt Classifier. citeseerx.ist.psu.edu

Index Terms
Computer Science      Pattern Recognition

Keywords
Sentence Boundary Detection   Verb Suffix   Abbreviation