International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 3, May 2014, 179-186

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

IMPROVEMENT OVER IL-POST TAGSET FOR KANNADA

BHUVANESHWARI C. MELINAMATH

Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India

ABSTRACT

Compilation of tag set is an important task in all NLP. It is the initial stage in all NLP applications. We are focusing on improvements over the IL-POST (Indian language part of speech tag set) in this paper. Our tag set is fine grained and captures detail information. We have developed this tag set keeping higher NLP applications in mind. Fine grained tag set is useful for NLP applications like chunking, parsing, morphological analyzer and machine translation etc. We follow EAGLES (Expert Advisory Group on Language Engineering Standards) as guideline with modifications as required for our Kannada Language.

The morphology of Kannada is complex as comparable to Turkish and Finnish. This tag set can be adopted for whole Dravidian language family. This is Hierarchical tagset and is largely based on computational needs. We have compiled a tag set of 170 tags. Compilation of tag set is an important task in all NLP and is quite challenging for Languages like Kannada. This paper will look at solving the open issues left unsolved in Microsoft's IL-POST tag set like clitics, auxiliaries. Modal auxiliaries etc. Tagging efficiency rate is more than 90% in Our tag set as compared existing ones.

KEYWORDS: Expert Advisory Group on Language Engineering Standards) EAGLES, Machine Translation (MT), Natural Language Processing (NLP), Part of Speech (POS)