Hierarchal POS tagging for Tamil language using Machine learning approach

[1]Dr.V.Dhanalakshmi, HOD, Department of Tamil, FSH, SRM University, Kattankulathur. (dhanagiri@gmail.com).

[2]Mr.Anand Kumar, Assistant Professor, CIET, Coimbatore. (mailtoanandkumar@gmail.com)

Abstract:

This paper presents the intricacies involved in developing a hierarchal **POS** tagger generator using SVMTool for Tamil language. Tamil, a Dravidian language has a very rich morphological structure which is agglutinative. Tamil words are made up of lexical roots followed by one or more affixes, mostly suffixes. So tagging a word in a language like Tamil is very complex. We try to resolve this complexity by identifying the categorical ambiguities and developing three hierarchal tag sets at word grammatical category and grammatical feature level. These tag sets were used to annotate the corpora and trained using the SVMTool (An Open source tool available at http://www.lsi.upc.es/~nlp/SVMTool) to generate the POS tagger model. The results obtained in each level were encouraging.

References:

1. Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai. 2006. AnnCorra:*Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages*, Technical Report, Language Technologies Research Centre IIIT, Hyderabad.
2. Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S,2008, *Tamil Part-of-Speech tagger based on SVMTool*, Proceedings of the COLIPS International Conference on Asian Language Processing 2008 (IALP), Chiang Mai, Thailand. 2008: 59-64.
3. Dhanalakshmi V, Anandkumar M, Shivapratap G, Soman, K P, Rajendran S. May 2009. *Tamil POS Tagging using Linear Programming,* In International Journal of Recent Trends in Engineering, Vol:1(2):166-169.
4. Gim´enez, J and L M`arquez, 2003. *Fast and Accurate Part of- Speech Tagging: The SVM Approach* Revisited, in Proceedings of the Fourth RANLP.