==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 15:11 November 2015
==

A Coherent Scrutinization on Syntactic Categories for Tagging Tamil Lexicon

Dr. (Mrs.) Ananthi Sheshasaayee, MCA., M.Phil., Ph.D.
Angela Deepa.V.R., M.Sc., B.Ed.

==

Abstract

The arrangement of words based on rules is termed as Syntax. Natural languages have their renowned syntactic rules that demonstrate their latent features. It is attributed in a form of free word order and some have conditions on the word order arrangement. As a consequence, the smallest unit in a sentence called word or lexicon has its unique function which determines the nature of the sentence. The categorized groups of functionalities of the words are termed as syntactic categories. The syntactic categories are also termed as Parts of Speech. Numerous NLP application benefits from this syntactic information, but for morphological rich languages like Tamil, the problem of tagging the every word in a particular part of speech remain a exigent task. This paper reports about the various approaches used for developing POS tagging and the developed POS taggers particularly for the Tamil language is discussed.

Keywords: Tag Set, Suffix, Prefix, Parts-of-Speech, Tagging, Morphological Analysis, Hidden Markov Model (Hmm).

Introduction

The importance of parts-of-speech for language processing is about the detailed information it gives to the word and their neighbors. It is also termed as POS, word classes, morphological classes and lexical tags. The computational methods used in assigning parts-of-speech categories of words are termed as parts-of-speech tagging. Syntactic categories or parts-of-Speech tagging is defined as the process of marking the word [1] in a text in a particular part of speech according to a context. This plays a predominant role and serves as a preprocessing step in most of the NLP applications like information retrieval, Word disambiguation, Speech recognition, Machine translation, Name entity recognition, Text to speech, etc. Since numerous NLP applications rely on the syntactic categorical information, the need for developing an efficient POS tagging is important. Although the tagging of Indian languages gained interest in recent times the usage of tag sets by different research scholars leads to a chaotic situation. Standardization is the only dimension that can solve this discrepancy. Dravidian languages like Tamil are morphological rich in content and agglutinative in grammatical nature. Deep analysis is required at appropriate levels [1] to understand the feature of the languages. `

Taxonomy of POS Tagging

POS tagging is broadly classified into supervised and unsupervised tagging. The following (figure. 1) demonstrates the different POS tagging used for natural language texts [2] linguistic rule, stochastic and a combination of both. Supervised tagging is a method that helps the system to learn the rules of tagging. It is based on pre-tagged corpus. Unsupervised tagging is an alternative method that uses algorithms to tag automatically the tag sets. It does not require a pre-tagged corpus. They are further divided into two distinct approaches for POS Tagging-Rule based and Stochastic approach [1]. The use of large database which consists of manual written linguistic rules to order the morphemes and the relative contextual information is termed as Rule based approach. The Stochastic approach involves the usage of unambiguously tagged text, which estimates the probabilities in selecting the most likely sequence.

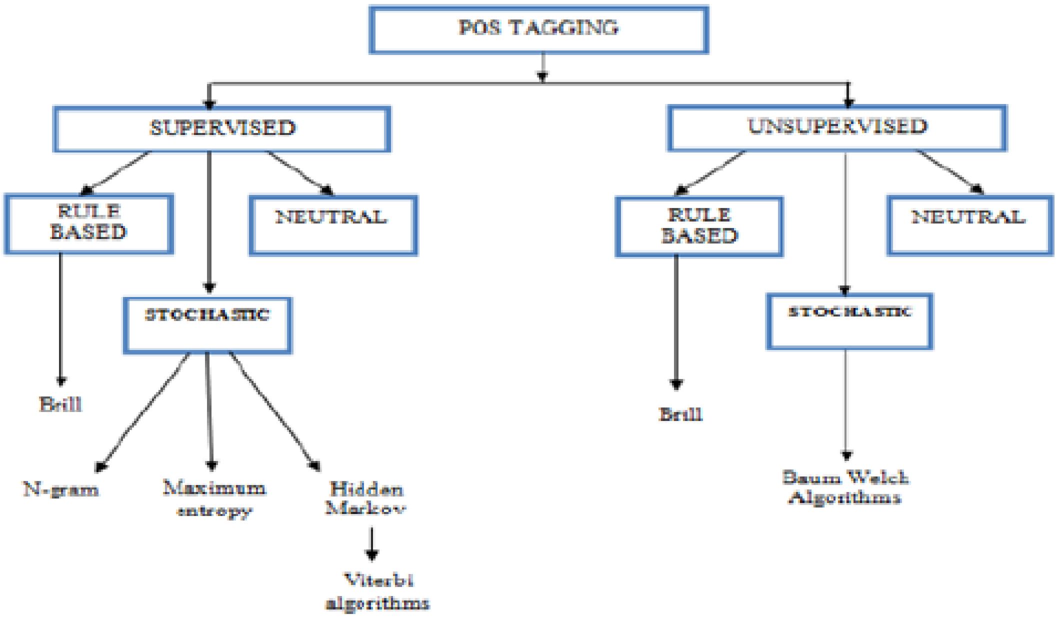

Figure 1: Various approaches of POS tagging

Tagging Methodologies

Tnt-A Statistical Part-of-speech tagger

TnT tagger is a Hidden Markov Model, using the second-order form involving supervised tagging technique. This statistical model [7] consists of states representing tags in which the output word is predicted. The transition probabilities mainly rely on the pair of tags in which the output probabilities depend on the current category. This transition and output probabilities are based on the tagged corpus. Based on the relative frequencies, maximum likelihood probabilities were calculated. The contextual and lexical frequencies of words are handled relatively according to their presence in the lexicon.

Shallow parsing with conditional random fields

Conditional random fields (CRFs) is a probabilistic framework [3] for segmenting, labelling sequences, trees and lattices. The conditional probability distribution gives a particular observation sequence over the label sequence rather than a joint distribution over both label and observation sequences. The predominant [8] feature of CRFs is that their conditional nature. It outperforms the HMM in the factor of relaxation of the independent assumptions which is the basic requirement for the HMM to ensure tractable inference. CRFs inhibit the bias caused by labelling. As a result, CRFs claims for the better result from the field of Bioinformatics, computational linguistics and speech recognition when compared to the maximum entropy Markov model (MEMMs) exhibiting labelling and the conditional Markov model of directed graphical models.

Maximum entropy model

Maximum Entropy Model is a probability framework [9] as it involves linguistic modelling. This model involves a flexible technique which is defined as H*T where H set of possible words, the context of the tags and T is the set of allowable tags. For a given training data, sequence of words {w1,….,wn} and tags{t1,…tn} based on the available histories the unseen word is predicted. The relevant parameters are chosen based on the maximize likelihood of the training data. Thus the POS tags are predicted simultaneously based on many contextual features.

A simple rule based tagger

The rule based parts-of-speech tagger tends to automatically acquire the rules and tags. This simple tagger doesn’t need a large storage allocation. It predominantly relies on the [10] reasonable, meaningful rules. Rule based tagger can be deployed on any corpus with different tag sets. These taggers efficiently learn the rules which encourage the researchers for a creation better rule template. Recognizing and remedying from the learnt rules the tagger incrementally improves its accuracy.

SVM tool

The SVM tool is based on Support Vector Machines. This software package consists of three main parts. Initially, the learner also called as SVM learner followed by the tagger termed as SVM tagger proceeds by an evaluator named as SVM teval. These SVM models [11] learn the component through the SVM learner upon the Training corpus. At the time of tagging appropriate strategy can be utilized, which can suit the need and purpose of the scenario. Initially, the SVM teval component displays the output based on the trained corpus to that of the predicted annotation of the SVM tool. This tool is an effective, efficient and flexible parts-of-speech tagger that perfectly fits into any environment. Adaptability Features of the SVM tool perfectly blend to the configuration that can match accurately to build most of the applications.

Tamil Language

Dravidian languages are spoken by more than 250 million people, primarily residing across the southern parts of India. There are four major languages Kannada, Malayalam, Tamil and Telugu [12]. Among them Tamil language is considered to be the longest surviving language. This language stands for its rich literary tradition. It consists of three periods [2] of literature values categorized as Old Tamil (300 BCE-700CE), Middle Tamil (700-1600) and Modern Tamil (1600-present). From linguistic part of view Tamil language has verb final and a free word order form to some extent.

Challenges Faced in Understanding the Structure of Language

Tamil is a classical language that is spoken by more than 66 million people all over the world. It is a morphological rich content and complex in structural notes. It inflects to person, gender and invariably joins with auxiliaries to indicate mood, aspect, attitudes, etc. It takes both lexical and inflectional morphology [13]. The noun form of the language takes up almost five hundred word forms inflecting based on the post positions. The verb form inflects with the auxiliaries taking up approximately two thousand word forms. To incorporate this language in the machine understandable mode the words are needed to be tagged at the root level to build a successful language application.

POS Tagging in Tamil

Labeling a part of speech or lexical case marker for every word in a sentence to denote its grammatical notions is termed as parts of speech (POS) tagging. This process is analogous to the tokenization process of understanding computer languages. Dravidian languages like Tamil which are morphological rich in content and agglutinative in nature faces a complex issue in tagging their word content. Almost all lexemes in Tamil language incorporate bounded morphemes which are structured as inflectional and derivational morphology. High rate of complexity is faced in tagging Tamil lexicons. Ambiguities in words and to resolve the complexity of Tamil lexicons deep analysis have to be done, which remains a demanding chore for Tamil POS tagging.

Examples: **Noun (+ number) (+ case)**

Paravai-kaL-ai <NN>

'Birds' -plural-accusative case suffix

Verb stem (+ Tense) + (Person-Number) + (Gender)

Nata +nt +En

'I walked'

Parts of Speech Taggers for Tamil Language

Many approaches have been used to build an efficient Parts-of-speech tagger for the Tamil language. The following are the various methodologies deployed in building a POS tagger.

Rule based POS tagger

This rule-based POS tagger targets the major tags and the sub tags based on the meaningful rules. A hybrid POS tagger using HMM [14] techniques is also used for tagging. The out frame of these models can be divided into three stages: Pre-editing, automatic tag assignment and manual post-editing. The corpus is converted into a suitable format based on the POS tags during the pre-editing stage. After the initial assignments, the words are manually corrected for accuracy.

Projection and Induction Techniques

Well defined morphological rules are used to build this POS tagger for the Tamil language. Projection and induction techniques are used in these systems. Application of alignment and projection techniques for the process of constructing POS tag incorporates the induced root words from English-Tamil alignment, lemmatization followed by morphological induction techniques. In the experiments, rule based morphological [4] analyzer and POS tagger were built with 85.56% accuracy. Based on the English via alignment across a sentence aligned corpora the required categorical information and root words are obtained. Improvements are made in rule based morphological analyzer and POS tagger based on the categorical information; root words are obtained from POS projection and morphological induction respectively via sentence aligned corpora. This method generated nearly 600 POS tags in the POS tagging process.

Lexical phonological approach

The tagTamil [14] is based on this lexical phonological approach. Tagging and generation are done by these tagTamil. It targets the verb forms by morphotactics, by morphologically processing them using an index method.

Morpheme based language model

This model [16] consists of information about the stem type, last morpheme and previous to the last given morpheme. According to the information gathered in the language model categorization of part of speech is developed. The Generative scaling technique is used for estimation. It deems to the stem word along with their lexical category and morphological components. Output is the best tag for each word based on the strings of words and tag sets given as input.

Hierarchical POS tagging

This tagging method [5] identifies the categorical ambiguities by deploying a hierarchical structure for the tag set at the grammatical category and feature level of the words. SVM tools were used to annotate the corpora to generate this POS tagger model.

Conclusion

Tamil is a morphological complex structure which needs deep feature understanding to integrate it in a machine understandable form. Morphological analysis is needed for complex structured languages. Parts-of-speech tagging is an important pre-processing step in the morphological analysis. The Parts-of-speech tagging plays a vital role in building any Natural language processing applications. Therefore, from this study, it is found that due to the less availability of resources, developing an efficient tagger for Tamil language remains a challenging task. Though a considerable amount of machine learning approaches is used in developing POS taggers, there is no remarkable development for Tamil POS tagging. Therefore, different approaches can be used in the future for building efficient POS tagging.

==

References

[1] Ananthi Sheshasaayee and Angela Deepa. V. R, "The Role of Morphological Analyzer and Generator for Tamil Language in Machine Translation Systems", International Journal of Computer Sciences and Engineering, Volume-02, Issue-05, Page No (107-111), May -2014

[2] Kumar, Dinesh, and Gurpreet Singh Josan. "Part of speech taggers for morphologically rich Indian languages: a survey." International Journal of Computer Applications 6.5 (2010): 32-41.

[3] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.", 2001.

[4] Selvam, M., and A. M. Natarajan. "Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques." International journal of computers 3.4, 2009

[5] Dr.V.Dhanalakshmi and Mr.Anand Kumar, “Hierarchal POS tagging for Tamil language using Machine learning approach”, Unpublished.

[6] Jurafsky D and Marting J H, Speech and Language Processing An Introduction to Natural Language Processing,Computational Linguistics and Speech Recognition, Pearson Education Series 2002

[7] Brants, Thorsten. "TnT: a statistical part-of-speech tagger." Proceedings of the sixth conference on Applied natural language processing. Association for Computational Linguistics, 2000.

[8] Sha, Fei, and Fernando Pereira. "Shallow parsing with conditional random fields." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[9] Ratnaparkhi, Adwait. "A maximum entropy model for parts-of-speech tagging." Proceedings of the conference on empirical methods in natural language processing. Vol. 1. 1996.

[10] Brill, Eric. "A simple rule-based part of speech tagger." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.

[11] Giménez, Jesús, and Lluis Marquez. "SVMTool: A general POS tagger generator based on Support Vector Machines." In Proceedings of the 4th International Conference on Language Resources and Evaluation. 2004.

[12] A Comprehensive Survey on Parts of Speech Tagging Approaches in Dravidian Languages Proceedings of 09th IRF International Conference, 27th July-2014, Bengaluru, India, ISBN: 978-93-84209-40-72

[13] Dr. A.G. Menon, S. Saravanan, R. Loganathan Dr. K. Soman,, "Amrita Morph Analyzer and Generator for Tamil: A Rule Based Approach", Proceeding of Tamil Internet Conference, Coimbatore, India, Page no(239-243),2009.

[14] Arulmozhi P and Sobha L (2006), "A Hybrid POS Tagger for a Relatively Free Word Order Language", Proceedings of MSPIL-2006, Indian Institute of Technology, Bombay

[15] Vasu Renganathan,(2001),"Development of Part-of-Speech Tagger for Tamil", Tamil Internet 2001 conference.

[16] Pandian, S. Lakshmana, and T. V. Geetha. "Morpheme based Language Model for Tamil Part-of-Speech Tagging." Webpage http://www. gelbukh. com/polibits/38_02. pdf (2008).

Dr. (Mrs.) Ananthi Sheshasaayee, MCA, M.Phil., Ph.D.
Research Supervisor
PG&Research Department of Computer Science
Quaid-E- Millath Government College for Women
Chennai -600002
Tamilnadu
India
ananthi.research@gmail.com

Angela Deepa.V.R., M.Sc., B.Ed.
Research Scholar
PG&Research Department of Computer Science
Quaid-E- Millath Government College for Women
Chennai -600002
Tamilnadu
India
angelrajan.research@gmail.com