A COMPREHENSIVE SURVEY ON PARTS OF SPEECH TAGGING APPROACHES IN DRAVIDIAN LANGUAGES

MERIN FRANCIS

M-Tech Student, Viswajyothi College of Engineering and Technology, Vazhakulam, Muvattupuzha
E-mail: merinfrancis@yahoo.com

Abstract- Parts of speech tagging is the act of assigning each word in a document a tag that corresponds to the meaning of the word in the particular context. It is important and act as a basic step in many natural language processing application from word sense disambiguation to speech recognition. Due to the variety in grammatical constructs and morphological differences, the approaches for tagging in different languages are widely varying. The theorotical approaches include supervised learning approaches as HMM based model, Maximum entropy model, SVM based taggers and CRF based taggers or unsupervised approaches as rule based taggers. The languages considered are morphologically rich Dravidian languages as Telugu, Malayalam, Tamil and Kannada. Different approaches are compared on their accuracy and analysis is done.

Keywords- Parts of speech tagging, Dravidian languages, Morphological analysis, Hidden Markov Model (HMM), Maximum Entropy (MaxEnt), Support Vector Machine (SVM), Conditional Random Fields (CRFs).

I. INTRODUCTION

Part-of-Speech (POS) tagging is the process of assigning a part of speech or lexical class marker to each word in corpus according to the context and is used in wide range of applications. It plays fundamental role in various Natural Language Processing (NLP) applications such as chunking, text to text conversion systems, speech recognition, information extraction, and word sense disambiguation etc. Word-free languages have relatively complex morphological structure of sentences than other languages, so POS tagging playa an important role in them. Besides other disambiguation and language specific issues, lack of open annotated corpora resources retards the research activities in Dravidian languages, although POS Tagging in such languages has gained an increased interest in recent times. As different research scholars use different tagsets and hence, standardization is another major issue.

In this survey paper, different parts of speech tagging approaches proposed in Dravidian languages are discussed. Dravidian languages are the family of some 70 languages spoken by more than 250 million people in primarily in South Asia, namely India, Pakistan and Srilanka. The Dravidian languages are divided into South, South-Central, Central, and North groups; these groups are further organized into 24 subgroups. Indian constitution recognises four major literary languages namely Telugu, Malayalam, Tamil and Kannada which are the primarily spoken in AndhraPradesh, Kerala, TamilNadu and Karnataka respectively.

Rest of the paper is organized as follows. Different tagging methodologies are discussed in general in section II. Section III to VI discuss on POS tagging done in different Dravidian languages comparing the accuracy of various approaches. There are POS taggers based on SVM model, HMM model and MBLP model proposed in Malayalam. A CRF based approach, kernel based approach and probabilistic classifiers are proposed in Kannada.

A rule based approach is proposed in Tamil and Telugu. Conditional Random Fields and Transformation Based Learning is discussed in Telugu. In Tamil, a morpheme based approach and hierarchical tagging are also discussed. The conclusion and analysis is done in section VII.

II. TAGGING METHODOLOGIES

A. TnT-A Statistical part-of-speech Tagger

TnT tagger is a supervised tagging technique proposed in. It is a Hidden Markov Model based efficient statistical parts-of-speech tagger. Second order Markov model is used here. The states of the model represent tags, outputs represent the words in which transition probabilities depend on the pairs of tags and the output probabilities only depend on the most recent category. A tagged corpus is used to determine the transition and output probabilities. Initially, based on the relative frequencies, maximum likelihood probabilities are calculated and it is assumed to be zero if the corresponding numerators and denominators are zero. Smoothing of the contextual and lexical frequencies are done later by handling the words that are not in the lexicon. For the Penn Treebank, the authors claim an accuracy of 96.7%.

B. Shallow Parsing with Conditional Random Fields

Lafferty and et al. define conditional random fields (CRFs) as a probabilistic framework for labelling and segmenting structured data, such as sequences, trees

and lattices. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The authors show here how to train a conditional random field to achieve performance as good as any reported base noun-phrase chunking method on the CoNLL task. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models. CRFs outperform both MEMMs and HMMs on a number of real-world tasks in many fields, including bioinformatics, computational linguistics and speech recognition.

C. Maximum Entropy Model

The Maximum Entropy model is an extremely flexible technique for linguistic modelling, since it can use a virtually unrestricted and rich feature set in the framework of a probability model.

The probability model is defined over H*T where H is the set of possible word and tag contexts, or "histories", and T is the set of allowable tags. Given a sequence of words { w l , . . . , wn} and tags {tl,...tn} as training data, hi is defined as the history available when predicting ti. The parameters {p, a l . . . . . ak} are then chosen to maximize the likelihood of the training data. The model can be classified as a Maximum Entropy model and simultaneously uses many contextual "features" to predict the POS tag. The implementation claims an accuracy of 96.6% on unseen text.

D. A Simple rule based tagger

A simple rule-based part of speech tagger which automatically acquires its rules and tags with accuracy comparable to stochastic taggers is proposed in. The rule-based tagger has many advantages over these taggers, including: a vast reduction in stored information required the perspicuity of a small set of meaningful rules, ease of finding and implementing improvements to the tagger, and better portability from one tag set, corpus genre or language to another. The fact that a simple rule-based tagger that automatically learns its rules can perform so well should offer encouragement for researchers to further explore rule-based tagging, searching for a better and more expressive set of rule templates. The tagger works by automatically recognizing and remedying its weaknesses, thereby incrementally improving its performance. The authors claim to use a very simple algorithm with an error rate of about 7.9% when trained on 90% of the tagged Brown Corpus.

E. SVMTool

The SVMTool, is a simple, flexible, effective and efficient part–of–speech tagger based on Support Vector Machines. The SVMTool offers a fairly good balance among these properties which make it really practical for current NLP applications. It is very easy to use and easily configurable so as to perfectly fit the needs of a number of different applications. The SVMTool software package consists of three main components, namely the learner (SVMTlearn), the tagger (SVMTagger) and the evaluator (SVMTeval). Previous to the tagging, SVM models (weight vectors and biases) are learned from a training corpus using the SVMTlearn component. Different models are learned for the different strategies. Then, at tagging time, using the SVMTagger component, one may choose the tagging strategy that is most suitable for the purpose of the tagging. Finally, given a correctly annotated corpus, and the corresponding SVMTool predicted annotation, the SVMTeval component displays tagging results. The system claims to achieve an accuracy of 97.16% for English on the Wall Street Journal corpus.

III. MALAYALAM

Parts of Speech tagging approaches done in Malayalam are as follows.

B. HMM based Tagging

For Malayalam language, a stochastic Hidden Markov Model (HMM) based parts of speech tagger is suggested. For this purpose an annotated corpus is necessary. The authors claim 90.5% accuracy.

C. SVM based tagging

A machine learning approach with Support Vector Machine (SVM) is developed in. The authors claim 87% accuracy.

D. MBLP approach

The MBLP approach is based on the concept of Memory based Language processing (MBLP) algorithm. Two powerful techniques used in this approach are: the efficient storage of solved examples of the problem, and similarity based reasoning on the basis of these stored examples to solve new ones. TiMBL (Tilberg Memory based Language) tagger tool is used for the implementation. Testing is done with existing SVM-tagger for Malayalam POS tagging.

IV. KANNADA

Kannada is one of the South Indian languages with rich morphology. It is an ancient language which came into existence around 2000 years back, so it is declared as classical language. Approximately 6 crores of people speak Kannada around the world. The canonical word order of Kannada is SOV (subject–object–verb) as is the case with Dravidian

languages. Kannada is a highly inflected language with three genders (masculine, feminine, and neuter or common) and two numbers (singular and plural). It is inflected for gender, number and tense, among other things.

A. CRF Based Approach

Pallavi and et al. propose a POS tagger for Kannada using Conditional Random Fields (CRFs). POS tagger is being developed using Java programming and CRF++ (Yet Another CRF++ toolkit) of Linux platform. To train and test the system, the data is collected from on-line Kannada newspapers, and then manually processed and tagged. The proposed model is claimed to have good accuracy.

B. Kernel Based Approach

The paper in presents the development of a part-of-speech tagger for Kannada language that can be used for analyzing and annotating Kannada texts.. In order to alleviate problems for Kannada language, a new machine learning POS tagger approach is proposed. Identifying the ambiguities in Kannada lexical items is the challenging objective in the process of developing an efficient and accurate POS Tagger. Tagset of 30 tags was developed to build a part-of-speech Tagger using Support Vector Machine (SVM). A corpus of texts, extracted from Kannada newspapers and books, is manually morphologically analyzed and tagged using our developed tagset. The performance of the system is evaluated. The authors claim that the result obtained was more efficient and accurate compared with earlier methods for Kannada POS tagging.

C. Probabilistic Classifiers

Of the entire supervised machine learning classification algorithms, second order Hidden Markov Model (HMM) and Conditional Random Fields (CRF) is chosen in this work for POS tagging of Kannada language. Training data includes 51,269 words and test data consists of around 2932 tokens. Both set being disjoint and taken from EMILLE corpus. The authors claim that experiments show the accuracy of the tools based on HMM and CRF is 79.9% and 84.58% respectively.

V. TELUGU

Telugu language is spoken by 100 million native speakers in Andhra Pradesh. Telugu language occupied 15th position in the world and 2nd position in India. Telugu is a highly inflectional and agglutinative language providing one of the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms. Telugu generally follows the Subject Object Verb (SOV) structure.

A. Conditional Random Fields and Transformation Based Learning

Part Of Speech (POS) tagging and chunking using Conditional Random Fields (CRFs) and Transformation Based Learning (TBL) for Telugu, Hindi and Bengali is proposed in. Improved training methods based on the morphological information, contextual and the lexical rules (developed using TBL) were critical in achieving good results. The CRF and TBL based POS tagger has an accuracy of about 77.37%, 78.66%, and 76.08% for Telugu, Hindi and Bengali, and the chunker performs at 79.15%, 80.97% and 82.74% for Telugu, Hindi and Bengali respectively.

B. Rule based tagging

There are various functional modules which works together to give tagged Telugu text. The pre-edited Telugu text is given as input to tokenizer which separates input text into separate sentences and each sentence to words for doing tokenization. These words are than given to MA for analysis. The Morph-to-POS translator then converts morphological analysis into their corresponding tags using pattern rules. The disambiguation problem is handled by the POS disambiguator which reduces the problem of POS ambiguity. This ambiguity is reduced by unigram and bigram rules. Finally, the tagged text is produced by annotator.

VI. TAMIL

Tamil is one of the longest surviving classical languages in the world. Tamil grammar is agglutinative in nature. Suffixes are used to mark class, number and cases attached to a noun. Tamil word may have a lexical root to which one or more affixes are attached. Other features are plural for honorific noun, frequent echo words, and null subject feature i.e. all sentences do not have subject, verb and object.

A. Rule Based Morphological Analysis improved via Projection and Induction Techniques

Rule based morphological analyzer and POS tagger is built from well defined morphological rules of Tamil. They can be further improved by the root words induced from English to Tamil through the sequence of processes like alignment, lemmatization and induction with the help of sentence aligned corpora like Bible corpora, TV news, newspaper articles since finding the root in the inflected words is very difficult and leads to ambiguity. In the experiments, rule based morphological analyzer and POS tagger were built with 85.56% accuracy. POS tagged sentences in Tamil were obtained for the Bible corpus through alignment and projection techniques and categorical information had been obtained. Root words were induced from English to Tamil through alignment, lemmatization and induction processes. Further 7% improvement was made in rule based morphological analyzer and POS tagger using

categorical information and root words obtained from POS projection and morphological induction respectively via sentence aligned corpora

B. Morpheme based Language Model

A part-of-speech tagger for Tamil, which uses a specialized language model is proposed in A language model based on the information of the stem type, last morpheme, and previous to the last morpheme part of the word for categorizing its part of speech is developed. For estimating the contribution factors of the model, generalized iterative scaling technique is followed. The language model that considers the lexical category of the stem along with morphological components of a word are used to determine its POS tag. In case the word is equal to a stem then its POS category is the same as the stem lexical category. In case the word consists of a stem and a single morphological component then the language model is designed by considering both the lexical category of the stem and the morphological component. The input of a tagger is a string of words and a specified tag set similar to the described is done. The output is a single best tag for each word. The overall accuracy of this tagger is claimed to be 95.92%. The overall F-measure of 96% is claimed.

C. Hierarchal POS tagging

Intricacies involved in developing a hierarchal POS tagger generator using SVMTool for Tamil language is presented in the abstract of. By identifying the categorical ambiguities and developing three hierarchal tag sets at word grammatical category and grammatical feature level, the author's try to address the complexity of the agglutinative nature of Tamil language. These tag sets were used to annotate the corpora and trained using the SVMTool to generate the POS tagger model.

CONCLUSION

The Part of Speech tagging is the most important activity of any Natural Language based applications. The accuracy of any NLP tool is dependent on the accuracy of POS tagger. Different approaches have been used by authors for the development of part of speech tagger for Dravidian languages. Like many of the Indian languages, most of the Dravidian languages are resource poor. Even if resources exist somewhere, they exist without public access. From this study, it is found that the Dravidian languages are morphologically rich languages. So, morphological analyzer plays a vital role in developing a POS tagger. Further, machine learning based approaches gives somewhat better results as compared to other approaches. Very limited work has been done on Indian languages for part of speech tagging. So, different approaches can be used for the development of efficient tagger.

REFERENCES

[1]. Thorsten Brants, "TnT -- A Statistical Part-of-Speech Tagger.", Saarland University Computational Linguistics,Germany, 2000.

[2]. Fei Sha and Fernando Pereira, "Shallow Parsing with Conditional Random Fields ", 2003.

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

[4]. Eric Brill, "A Simple Rule-Based Part-of-Speech Tagger",1992.

[5]. Jes ´us Gim´enez and Llu´ıs M`arquez , "SVMTool: A general POS tagger generator based on Support Vector Machines ", TALP Research Center, LSI Department Universitat Polit`ecnica de Catalunya, 2004.

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

[7]. Jisha P Jayan and Rajeev R R, Parts Of Speech Tagger and Chunker for Malayalam –Statistical Approach, The International Institute for Science, Technology and Education (IISTE). ,Vol. 2, No. 3, December 2009.

[8]. Robert Jesuraj K, P. C. Reghu Raj, MBLP approach applied to POS tagging in Malayalam Language , NCILC 2013.

[9]. Anish A, Parts of speech tagging for Malayalam, Unpublished, July 2008.

[10]. Antony P.J, Santhanu P Mohan, Soman K.P, "SVM Based Part of Speech Tagger for Malayalam", IEEE International Conference on Recent Trends in Information,Telecommunication and Computing, pp. 339-341, 2010

[11]. Pallavi and Dr. Anitha S Pillai, Parts Of Speech (POS) Tagger for Kannada Using Conditional Random Fields (CRFs), NCILC 2014.

[12]. Antony, P. J., and K. P. Soman. "Kernel based part of speech tagger for Kannada." Machine Learning and Cybernetics (ICMLC), 2010 International Conference on. Vol. 4. IEEE, 2010.

[13]. BR, Shambhavi, and Ramakanth Kumar P. "Kannada Part-Of-Speech Tagging with Probabilistic Classifiers." International Journal of Computer Applications 48.17, 2012.

[14]. PVS, Avinesh, and G. Karthik. "Part-of-speech tagging and chunking using conditional random fields and transformation based learning." Shallow Parsing for South Asian Languages 21, 2007.

[15]. 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.

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

[17]. Lakshmana Pandian, S., and T. V. Geetha. "Morpheme based Language Model for Tamil Part-of-Speech Tagging." polibits 38, 2008

[18]. 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.

★★★