International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Parts of Speech (POS) Tagging in Telugu Corpora Using CRF Algorithm

Rajula Valaraju

Research Scholar, Department of Linguistics and Contemporary English, EFL University, Hyderabad, Telangana, India
Email: *valas.508[at]gmail.com*

Abstract: ***The study of NLP (Natural Language Processing), a branch of computer science and AI (Artificial Intelligence), enables machines to comprehend human language effectively and assist with linguistic tasks. The initial step in every NLP task is POS (Parts of Speech) tagging, which assigns a tag to a word based on its meaning and context. The present paper discusses parts of speech tagging (POS) in Telugu using Conditional Random Fields (CRF), a sequence modelling algorithm that is particularly effective in identifying entities or text patterns, such as POS tags, in highly inflectional and agglutinative languages like Telugu. Telugu is a highly inflectional and agglutinative language widely spoken in the southern part of India (mainly Andhra Pradesh and Telangana). The Language belongs to the Dravidian Family and, it follows the S - O - V structure. Compared to other machine learning algorithms, CRF has been proven more effective in overcoming label - bias problems in a language. In order to understand the language features and to tag the test corpus, an annotated corpus of 62, 996 words and a tag set of 18 tags is used for the study. The present study has achieved an accuracy of 80.17%.***

Keywords: POS tagging, CRF Model, BIS Tag set, Telugu Language

1. Introduction

A Part - of - Speech Tagger (POS Tagger) is a software tool that analyses text and assigns grammatical categories to each word, such as nouns, verbs, adjectives, etc. POS taggers are categorized into three types: rule - based, statistical, and hybrid. The rule - based taggers apply hand - written rules to assign tags to the given word in a context. However, rule - based taggers are non - automatic, very costly, and time - consuming. In contrast, the statistical taggers utilize the frequency or probability occurrences of a given word to assign the POS category. Finally, the hybrid POS tagger is a combination of both rule - based and statistical taggers. The performance of all the NLP tasks, such as grammar checker, parser, machine translation, etc., depends on the accuracy of the POS tagger, making it a crucial area of research and development in the field of NLP.

Telugu is a member of the Dravidian language family, spoken predominantly in the Indian states of Telangana and Andhra Pradesh as the first language and in Tamil Nadu and Karnataka as the second language. It is one of the 22 languages under schedule 8 of the constitution of India. In 2008, Telugu became a member of one of the six classical languages in India. It is a morphologically rich and complex language. Regarding word order, the canonical word order in Telugu is SOV (i. e., it is a head - final and left - branching language) (Krishnamurthi, 1985).

2. Literature Review

Part - of - speech (POS) tagging in Indian languages presents challenges due to the shortage of publicly available pre - tagged language corpora for most of these languages. Developing a POS tagger requires either a set of handwritten rules or a large annotated corpus. But, due to the absence of these rules and a large corpus, POS taggers for Indian languages are not readily available. Many people have worked to improve the accuracy of POS taggers for Indian languages. This section provides a comprehensive overview of the efforts made in POS tagging across multiple languages.

Pillai et al. (2014) developed a POS tagging strategy for Kannada that uses Conditional Random Fields (CRF). They obtained their information from online Kannada publications, which were manually processed and tagged. The AU - KBC tag set, which comprises 45 tags, was utilized. Their examination included a 1000 - word test. They applied a combination of manually tagged words for training and testing to create and assess their CRF - based model. In their experiment, they used CRF++, a toolkit for CRF - based modelling. Their training and testing data included a vocabulary of 1000 words. The results obtained by PERL scripting have a remarkable accuracy of 99.49%.

Joshi et. al (2013) used the IL - POST tag set to build an HMM tagger for Hindi. They used a dataset of 15, 200 sentences (equivalent to 358, 288 words) from the tourism domain to train the HMM tagger. The researchers used contextual information inside the text to identify accurate word combinations in order to address word ambiguity. They attained a 92% accuracy rate when the Hidden Markov Model (HMM) POS tagger was applied to test data.

Shambhavi et. al (2012) used the Maximum Entropy technique to research POS tagging for Kannada. A manual tagging effort of 5, 1267 words was undertaken to train the POS tagger. The tag set contains 25 unique tags, and the word data came from the EMILLE corpus (Enabling Minority Language Engineering). This attempt attained an accuracy of 81.6%.

Ekbal et al. (2007) describe the challenge of applying statistical Conditional Random Fields (CRFs) for Part - of - Speech (POS) tagging for Bengali. The POS tagger was created with a tag set of 26 POS tags built for Indian languages. A total of 72, 341 and 20, 000 - word forms were used to train and test the POS tagger, respectively. Experiments have shown that lexicon, named entity recognizer, and various word suffixes efficiently solve

concerns relating to unfamiliar terms, significantly improving the POS tagger's accuracy. A test result of a 90.3% accuracy rate highlights the effectiveness of the suggested CRF - based POS tagger.

Ganesh (2006) developed a three - step rule - based POS tagging system for Telugu. Initially, the study manually tagged numerous texts to create a thorough collection of 56 tags encompassing all features of Telugu speech. Following that, input text or a corpus was processed by a Telugu morphological analyzer, producing relevant results. Finally, the system used around 524 pre - defined morpho - syntactic rules to resolve data ambiguities.

Kumar et al. (2006) paper tried to use four different models, namely, Conditional Random Field (CRF), Maximum Entropy Model (MEM), Hidden Markov Model (HMM), and Memory - based learning. All these Machine Learning algorithms are trained on an annotated corpus of 27, 000 words, and then later they are tested on a dataset of 6, 000 words. The study shows a maximum accuracy of 82.5% was achieved on the testing data using the HMMs.

3. Collection of Corpus

For the purpose of the present study, we have compiled a collection of Telugu corpus from the Technology Development for Indian Languages (TDIL) corpus, which is the sample corpus of the Telugu mono - lingual text corpus. The TDIL corpus consists of Telugu articles in Unicode format, which are extracted from printed documents from various fields and sources. The collected text documents are from different domains, such as sports, entertainment, agriculture, health, and history. The collected and stored text documents from the Telugu corpus are combined to form a single Unicode text file. Finally, this Unicode text corpus is pre - processed and divided into the train (50701 tokens) and test (12295 tokens) data sets. The total size of the corpus is 62, 996 tokens.

For the present research, the tag set chosen is the BIS tag set framework, which aims to standardize morphosyntactic categories for all Indian languages. This tag set has been recommended for the standard tag set for annotating Indian languages. The tag set incorporates the advice of experts from NLP and language technology of Indian languages.

4. Implementation of the CRF Tagger

This section deals with the proposed POS tagger for the Telugu language using the CRF algorithm. Conditional random fields, or CRF, designed by Lafferty et al. in 2001, is a class of probabilistic discriminative models well suited for structured prediction tasks in which the contextual information or state of the neighbours influences the current prediction. According to Kumar et al. (2010), CRF is an undirected graphical model that defines a single long - linear distribution for a label sequence based on a specific observation sequence. To develop the POS tagger, we have downloaded the CRF tagger version of ++0.58 from online sources. The figure shows the proposed POS tagger for the Telugu language.

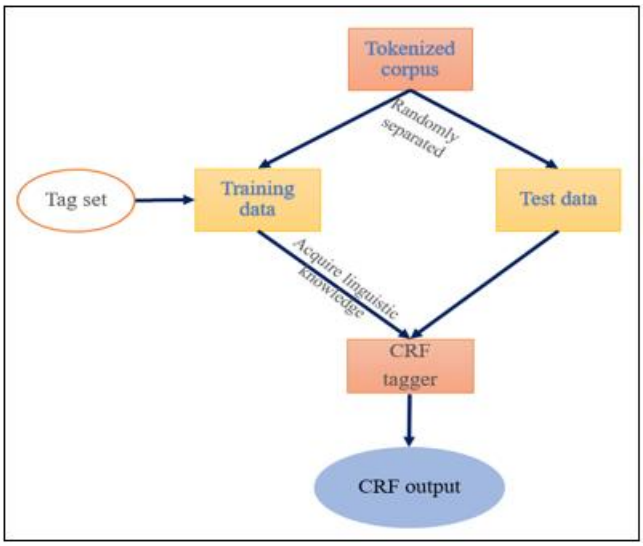

Figure: Proposed POS tagger

The two main components required for tagging in the CRF module are a test corpus and a labelled input training corpus. The CRF tagger acquires the lexical rules from the training corpus and automatically assigns a possible tag to the unlabelled test corpus. After training the model, the test or unannotated corpus is given as input to the CRF algorithm. Once the corpus has been trained and a model file created, the CRF tagger decodes this model using a test command. The trained model file and test data are inputs, and the tagged output is redirected to a different text file. The following screenshot displays the CRF tagger output for the Telugu language.

Screenshot1: CRF output

word	OriginalTag	Predicted Tag	
బాధలు	N_NN	N_NN	
,	RD_PUNC	RD_PUNC	
లక్షణాలు	N_NN	N_NN	
తగ్గిపోయినా	V_VM	N_NN	
కూడా	RP_RPD	RP_RPD	
వైద్యులు	N_NN	N_NN	
సూచించినంత		V_VM	N_NN
పూర్తికాలం	N_NN	V_VM	
మందులు	N_NN	N_NN	
మానకుండా	V_VM	V_VM	
వేసుకోండి	V_VM	V_VM	
.	RD_PUNC	RD_PUNC	
ఇలాంటి	JJ	JJ	
వారికి	PRON	PRON	
కేవలం	QT	QT	
శమన	N_NN	N_NN	
చికిత్సలే	N_NN	N_NN	
చేస్తారు	V_VM	V_VM	
.	RD_PUNC	RD_PUNC	
కొన్ని	QT	QT	
పశువులకు	N_NN	N_NN	
కృత్రిమ	N_NN	N_NN	
గర్భధారణ	N_NN	N_NN	
చేయించిన	V_VM	V_VM	
తర్వాత	N_NN	N_NN	
కూడా	RP_RPD	RP_RPD	
10	QT	QT	
గంటల	N_NN	N_NN	
వరకు	PSP	PSP	
ఎద	N_NN	N_NN	

5. Results and Analysis

The percentage of correctly tagged words determines the accuracy of the tagger. Though Telugu is an agglutinative and

morphologically rich language, the present study has achieved better results using the CRF algorithm. The accuracy achieved by the CRF tagger for the Telugu language is found to be 80.17%.

Apart from identifying and analysing the incorrect tags generated by the final tagged output of the CRF tagger, there are certain tags that remain unidentified by the CRF POS tagger. The total number of tokens in the test data comprises 12295 tokens, out of which 9858 tokens were found to be accurately tagged and 2437 were incorrectly tagged. It was observed that the errors were mostly dominated by proper nouns (NNP) and nouns (NN).

6. Conclusion

In this paper, we have developed a POS tagger for the Telugu language using the CRF algorithm. For the development, we have used 62, 996 tokens. The performance of the current approach is good, and the results obtained from the above approach were better. We believe that future improvements could enhance tagging accuracy by expanding the size of the tagged corpus.

References

[1] Brown, C. P. (1857). *A grammar of the Telugu language*. W. H. Allen and Company.
[2] Ekbal, A., Haque, R., & Bandyopadhyay, S. (2007, December). Bengali part of speech tagging using conditional random field. *In Proceedings of seventh international symposium on natural language processing* (SNLP2007) (pp.131 - 136).
[3] Krishnamurti, B. H. & Gwynn, J. P. L. (1985). *A grammar of modern Telugu*. Oxford University Press.
[4] Joshi, N., Darbari, H., & Mathur, I. (2013). HMM based POS tagger for Hindi. *Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC - 2013), 3 (6),* 341 - 349. http: //dx. doi. org/10.5121/csit.2013.3639
[5] Kudo, T. (2017). *CRF++ Toolkit* (Version CRF++ 0.58) [Computer Software].
[6] https: //taku910. github. io/crfpp/
[7] Kumar, G. K., Sudheer, K., & Avinesh, P. V. S. (2006). *Comparative study of various machine learning methods for Telugu part of speech tagging.*
[8] Lafferty, J., McCallum, A., & Pereira, F. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Columbia University.* http: //www.cs. columbia. edu/~jebara/6772/papers/crf. pdf
[9] Pallavi, A. S. P., & Pillai, A. (2014). Parts of speech (POS) tagger for Kannada using conditional random fields (CRFs). https: //ieeexplore. ieee. org/stamp/stamp. jsp?tp=&arnumber=8321833&tag=1
[10] Shambhavi, B. R., Kumar, R., & Revanth, G. (2012). A maximum entropy approach to Kannada part of speech tagging. *International Journal of Computer Applications*, 41 (13), 9 - 12. http: //dx. doi. org/10.5120/5600 - 7852

Author Profile

Rajula Valaraju is a research scholar in the Department of Linguistics and Contemporary English at The English and Foreign Languages University, Hyderabad. He is working on POS tagging for the Telugu Language. His Current research interests include Computational Linguistics.