*International Journal of Engineering & Technology*, *7 (4) (2018) 2418-2421*

SPC

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*
*doi: 10.14419/ijet.v7i4.14900*
***Research paper***

# A comparison of features for POS tagging in Kannada

**Shriya Atmakuri, Bhavya Shahi, Ashwath Rao B* and Muralikrishna SN**

Department of Computer Science and Engineering, Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, Karnataka, India - 576104

* *Corresponding author E-mail:ashwath.rao.b@gmail.com*

## Abstract

This paper proposes a system of part of speech tagging for the South Indian language Kannada using supervised machine learning. POS tagging is an important step in Natural Language Processing and has varied applications such as word sense disambiguation, natural language understanding etc. Based on extensive research into methods used for POS tagging, Conditional Random fields have been chosen as our algorithm. CRFs are used for sequence modeling in POS tagging, named entity recognition and as an alternative to Hidden Markov Models. Three very large corpora are used and their results are compared. The feature sets for all three corpora are also varied. The best method for the task is determined using these results.

***Keywords:*** *Conditional Random Field; Indian languages; Kannada; Natural Language Processing; POS tagging*

## 1. Introduction

Part-of-speech tagging is a fundamental task in Natural Language Processing and Computational Linguistics. A part-of-speech refers to the label attached to a subset of words in a language which have similar grammatical roles. Part of speech tags are frequently used as an important feature for other natural language processing tasks such as word-sense disambiguation, named entity recognition, information retrieval, and machine translation. As such, a fast part-of-speech tagger with high accuracy is an essential component of any languages NLP toolkit and lays the groundwork for further research in the field.

To conduct part-of-speech tagging, a list of part-of-speech tags must first be defined. There have been a number of different ways of categorizing words in a language over time.

Traditional systems generally defined very few part-of-speech tags. The Sanskrit grammarian Yaska defined only four categories in his 5th century BC work, Nirukta. These are nama which includes nouns and adjectives, akhyata or verb, upasarga, which is a pre-verb or prefix, and nipata or particle. The Greek grammarian Dionysius Thrax, defined eight categories (noun, verb, participle, article, pronoun, preposition, adverb, and conjunction) in his 2nd century BC work, The Art of Grammar.

However, modern day linguists recognise that these definitions are too general and simplified.Recent treebanks of English have adopted Part-of-Speech tagsets with many word classes. The Brown Corpus, one of the first English language corpora created for processing by a computer, use 87 tags. The Penn Treebank ues 45 tags and the C7 tagset created in 1997 uses 146.

A coarse tagset is useful, if the number of words is relatively low in the corpus whereas a finer tagset is useful to capture the nuances of words when the corpus is large enough to accommodate it. The choice of tagset is further complicated by the variance in different languages. For example, Japanese has three different classes of adjectives while English only has one. We use the Unified Parts of Speech (POS) Standard in Indian Languages which is discussed further below.

Part-of-Speech tags are very useful in Speech applications in Kannada. The same word (e.g. hathi) may be pronounced differently depending on the meaning(Part-of-Speech) it carries. The word is pronounced differently when it is a noun and when it is a verb. More or less, the same notion is applicable for all words when they have different Part-of-Speech. The determining and employing the Part-of-Speech of a word will result in better Speech synthesis accuracy and better accuracy in Speech recognition.

By knowing the Part-of-speech of a word one can also determine nature and number of morphemes that can be attached to the word. Part-of-speech tag will help in parsing, word-sense disambiguation algorithms and in shallow parsing to find names, times, dates or other named entities in the information extraction applications.

## 2. Related work

Antony P.J., Soman K.P. 2010 [1] present the development of a part-of-speech tagger for Kannada. The researchers have developed their own tagset consisting of 30 tags. The tagset comprises 5 tags for nouns, 1 tag for pronoun, 8 tags for verbs, 3 for punctuation, two for numbers and 1 each for adjective, adverb, conjunction, echo, reduplication, intensifier, postposition, emphasize, determiner, complementizer, and question word. They have built a part-of-speech tagger for Kannada using a Support Vector Machine (SVM). A dictionary of all words in the training data with all possible tags is considered. Binarizing of classification is done by assigning the correct tag a positive label, all other tags in the dictionary a negative label. Binarization of the classification is required as SVMs are binary classifiers. When trained on 10,000 words the accuracy is 48%, and as the training size increased to 54,000 words the accuracy increased to 86%.

Shambavi B R and Ramakanth Kumar [2] compare two different probabilistic models (HMM and CRF) on a portion of the EMILLE corpus. They utilise the tagset of 25 part-of-speech tags created by Bharati et al for the Indian Language Machine Translation Project. They use 95% of their data (51,269 words) for training only the remaining 5% (2,932 words) to test their models. They show that CRFs consistently outperform HMMs, achieving 84% accuracy with the former as opposed to 79% with the latter. While the comparison between CRFs and HMMs is useful for deciding between them in future work, the small size of their test dataset carries a risk that their model may be overfitted.

M. C. Padma and R. J. Pratibha [3] take an unusual approach in creating a part-of-speech tagger for Kannada using no probabilistic models. Their system uses morphological analysis along with lookup tables and a rule-based approach. Their choice of tagset is the BIS Dravidian tagset which is a hierarchical system with 11 categories at the top level. They test their model on four smaller datasets (with 1352, 892, 357, and 257 tokens respectively) which were extracted from the EMILLE corpus.They obtain an average precision of 88.75%. While their system is effective, a non-probabilistic approach is intensive in both effort and data. It is also inflexible to the evolution of the language.

K. P. Pallavi and Anitha S. Pillai [4] also use CRFs for POS tagging. They develop a tagger using an 80,000-word corpus created from Kannada Wikipedia. Their tagger achieves a maximum precision of 92.4% during cross validation.

CRFs have also shown their efficacy with other Indian languages. Notably, Avinesh and Karthik G [5] use Conditional Random Fields and Transformation Based Learning for part-of-speech tagging and chunking in Telugu, Hindi, and Bengali. They achieve an accuracy of about 77.37%, 78.66%, and 76.08% for the three languages respectively.

This paper expands on previous work by utilizing CRFs due to their proven efficacy and testing them on a much larger corpus than has been previously used. It also compares the effectiveness of a number of different features to obtain the optimal set of features for this task.

# 3. Methodology

This section describes the approach utilised in undertaking this work. It first expounds the idea behind Conditional Random Fields, which are a core facet of the approach. It then delves into the dataset used and its features. Finally, it outlines the various features designed and the variations attempted.

## 3.1. Conditional Random Fields

For the purpose of POS Tagging, Conditional Random Fields are used. CRFs, which were first described in [6] are a discriminative probabilistic model used to segment and label sequenced data. Unlike HMMs, which are generative, CRFs do not rely on the assumption of label independence. The advantage of discriminative models over generative ones is that they do not model the distribution of the features and instead focus is on modeling the distribution of labels over data. Another advantage of CRFs is that they allow the incorporation of data-dependent global features into the model which can be hard to do with generative models. Thus, CRFs are frequently used for Natural Language tasks such as POS tagging, named entity recognition and syntactic disambiguation.

CRFs assign a sequence of labels (in this case, the part-of-speech tags) with the highest probability to a sequence of inputs (in this case, the Kannada words). They are a supervised learning model which learn a set of feature functions and their corresponding weights to perform classification. CRFs can also be understood as an extension of logistic regression with structured output. The probability distribution for CRFs is modeled by:

$$p(y|x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t) \tag{1}$$

where $y$ is the sequence of labels, $x$ is the sequence of inputs $\Lambda = \lambda_k \varepsilon R$ is the parameter vector, $f_k$ is a set of real-valued feature functions and $Z(x)$ is the normalization function given by:

$$Z(x) = \sum_{y} \exp \sum_{k=1}^{k} \lambda_k f_k(y_t, y_{t-1}, x_t) \tag{2}$$

## 3.2. Dataset

We use the Kannada treebank developed in [7] to train our tagger. The treebank is divided into three corpora by topic. One corpus contains conversational data, the second contains tourism data, and the third is a general corpus. The general corpus is the largest and contains 17,175 sentences with a total of 218,530 tokens. The tourism corpus and the conversational corpus are around the same size and have 1,883 sentences with 26,521 tokens and 2,260 sentences with 19,315 tokens respectively. Each corpus was split with a 70-30 ratio for training and testing the POS tagger.

All tokens in the corpora are annotated with a POS tag, chunk, morphological information (including list of suffixes), dependency relations across chunks, sentence type, voice type. Of these features, the part-of-speech tag and the list of suffixes are significant in our task.

The corpora were tagged using the Unified Parts of Speech (POS) Standard in Indian Languages [8] which has been drafted by the Department of Information Technology, Govt. of India. This standard details the uses of labels for POS tagging and includes XML schemas for the common POS format. Our dataset uses 39 different tags for Kannada that have been detailed in the standard. The count of each of these tags in the dataset, along with their descriptions is shown in Table 1.

## 3.3. Models

Five different CRF models were tested using different template files. First, a very rudimentary model was built that used only the previous and current words as features for the tagger. The window size was then expanded to take into account the previous three words, the current word and the next three words. To contrast the effect of increasing the window size against using the inherent features present in the tokens, another model was trained using the previous, current and next words, along with a length feature. The length of the word was turned into a binary feature, with a value of 1 if the words had more than three characters and 0 otherwise. As this was more effective than the increased window size, further investigation was conducted into using the tokens inherent features.

Our penultimate model used the following features:

1. Context: The previous three words, the word to be tagged and the next three words.
2. Length: A binary feature with a value of 1 if the word was longer than three characters and zero otherwise.
3. IsDigit: A binary feature with a value of 1 if the token contained a digit and 0 otherwise.
4. IsPunct: A binary feature with a value of 1 if the token contained a non-alphanumeric character and 0 otherwise.
5. Suffixes: All the suffixes of the word.

**Table 1:** Distribution of POS Tags in the Dataset

| **Tag** | **Meaning** | **Count in Dataset** |
|---|---|---|
| CC_CCD | Conjunction (Co-ordinator) | 6772 |
| CC_CCS | Conjunction (Subordinator) | 3700 |
| CC_CCS_UT | Conjunction (Quotative) | 216 |
| CL | Particle (Classifier) | 16 |
| DM_DMD | Demonstrative (Deictic) | 3849 |
| DM_DMI | Demonstrative (Indefinite) | 561 |
| DM_DMQ | Demonstrative (Wh-word) | 360 |
| DM_DMR | Demonstrative (Relative) | 1 |
| JJ | Adjective | 10660 |
| NN | Noun | 7 |
| N_NN | Noun (Common) | 90220 |
| N_NNP | Noun (Proper) | 13182 |
| N_NNV | Noun (Verbal) | 451 |
| N_NST | Noun (Nloc) | 2954 |
| PR_PRC | Pronoun (Reciprocal) | 15 |
| PR_PRF | Pronoun (Reflexive) | 904 |
| PR_PRI | Pronoun (Indefinite) | 241 |
| PR_PRP | Pronoun (Personal) | 9830 |
| PR_PRQ | Pronoun (Wh-word) | 755 |
| PSP | Postposition | 720 |
| QT_QTC | Quantifiers (Cardinals) | 7244 |
| QT_QTF | Quantifiers (General) | 1208 |
| QT_QTO | Quantifiers (Ordinals) | 436 |
| RB | Adverb | 4474 |
| RD_ECH | Residuals (Echowords) | 57 |
| RD_PUNC | Residuals (Punctuation) | 30404 |
| RD_SYM | Residuals (Symbol) | 3481 |
| RD_UNK | Residuals (Unknown) | 1 |
| RP_CL | Particle (Classifier) | 5 |
| RP_INJ | Particle (Injection) | 166 |
| RP_INTF | Particle (Intensifier) | 601 |
| RP_NEG | Particle (Negation) | 153 |
| RP_RPD | Particles (Default) | 2777 |
| V_VAUX | Verb (Auxiliary) | 355 |
| V_VM | Verb (Main) | 45 |
| V_VM_VF | Verb (Finite) | 23849 |
| VM_VM_VINF | Verb (Infinitive) | 1391 |
| V_VM_VNF | Verb (Non-Finite) | 19986 |
| V_VM_VNG | Verb (Gerund) | 939 |

Suffixes are an especially important feature as Kannada, like all other Dravidian languages, is an agglutinative language. Words in Kannada contain many suffixes which modify their meaning and role in the sentence. For our final model, we varied the number of suffixes to observe their effect on the results and found that inclusion of only the first three suffixes as features is sufficient to produce accuracies comparable to the model that uses all suffixes.

## 4. Results

Table 2 summarizes the results achieved using the various models. The accuracies are calculated as follows:

$$Accuracy = \frac{No.ofcorrectlytaggedwords}{Totalno.ofwords} \tag{3}$$

**Table 2:** Results of Each Model

| | Conversational | Tourism | General |
|---|---|---|---|
| Previous Word | 71.6 | 75.6 | 80.5 |
| Window Size 3 | 78.1 | 78.5 | 83.6 |
| Length | 79.2 | 79.1 | 84.7 |
| Suffix | 83.2 | 83.7 | 89.1 |
| First 3 Suffixes | 83.5 | 83.1 | 89.1 |

The accuracy increases with each subsequent model, across all corpora. The highest accuracy achieved is by using the suffixes and other inherent features of the token. The General corpus achieves a higher accuracy than the other corpora due to its larger size. Since the size of the feature vectors used in the suffix method is very large, another method was tested which took only the first three suffixes of the token into account. The accuracy achieved through this method was comparable to the previous method, even though this method used less data, and hence was faster to train.

**Table 3:** Results of Training on the General Corpus

| Tested On: | Conversational | Tourism |
|---|---|---|
| Previous Word | 72.7 | 71.4 |
| Window Size 3 | 80.8 | 76.1 |
| Length | 81.2 | 76.5 |
| First 3 Suffixes | 85.7 | 79.4 |

**Table 4:** Results of Training on the Tourism Corpus

| Tested On: | Conversational | General |
|---|---|---|
| Previous Word | 42.0 | 47.7 |
| Window Size 3 | 57.7 | 63.6 |
| Length | 57.4 | 67.7 |
| First 3 Suffixes | 63.7 | 70.2 |

Table 3, Table 4, and Table 5 show the accuracies achieved when training the model on a particular corpus and testing the model on the other corpora. In general, the accuracies increase with each subsequent model. The General corpus gives the best accuracy on testing with other corpora because of its comparatively larger size. While the previous word model is not very robust across corpora, the other models which use more contextual information give fairly good results even when tested on a different corpus than their training corpus.

**Table 5:** Results of Training on the Conversational Corpus

| Tested On: | General | Tourism |
|---|---|---|
| Previous Word | 44.0 | 42.5 |
| Window Size 3 | 61.9 | 61.0 |
| Length | 67.6 | 64.4 |
| First 3 Suffixes | 69.1 | 62.7 |

## 5. Conclusion

In this paper, we develop a CRF-based part-of-speech tagger for Kannada. We test different combinations of features to define the ideal feature set and verify the robustness of the models by testing them on corpora which differ widely in topic and style. We also demonstrate that the size of the corpus has a clear and unmistakable effect on the performance of the tagger. We achieve a maximum accuracy of 89.1% on our largest corpus.

An effective POS tagger an lay the foundation for future work in the language such as chunking, or creation of a parse tree. Further work in improving the POS tagger might also be done using neural networks.

## Acknowledgements

We are thankful to Department of Electronics and Information Technology, New Delhi for funding the Treebank project. Also, thanks to Prof. Dr. Dipti Mishra Sharma, Professor and Head, Language Technologies and Research Centre (LTRC), International Institute of Information Technology, Hyderabad and her colleagues and students for their guidance and help in creating the Kannada treebank corpus.

## References

[1] PJ Antony and KP Soman. Kernel based part of speech tagger for kannada. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 4, pages 2139–2144. IEEE, 2010.

[2] Shambhavi BR and Ramakanth Kumar. Kannada part-of-speech tagging with probabilistic classifiers. *international journal of computer applications*, 48(17):26–30, 2012.

[3] MC Padma and RJ Prathibha. Morpheme based parts of speech tagger for kannada language. *World Academy of Science, Engineering and Technology, International Journal of Cognitive and Language Sciences*, 3(6), 2016.

[4] KP Pallavi and Anitha S Pillai. Kannpos-kannada parts of speech tagger using conditional random fields. In *Emerging Research in Computing, Information, Communication and Applications*, pages 479–491. Springer, 2016.

[5] Avinesh PVS 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.

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

[7] Ashwath Rao, SN Muralikrishna, and Ashalatha Nayak. Developing a dependency treebank for kannada. 2014.

[8] Govt. of India Department of Information Technology, Ministry of Communications & Information Technology. Unified parts of speech (pos) standard in indian languages.