Implementation on parts of speech tagging of Kannada language text

Saritha Shetty[1*], Sarika Hegde[2], Savitha Shetty[2*]

[1]Department of master of computer applications, Nitte deemed to be university, Nitte, India,574110
[2]Department of computer science and engineering, Nitte deemed to be university, Nitte, India,574110

*Corresponding author
E-mail : shettysaritha1@gmail.com, hegdesarika@gmail.com, shettysavi1@gmail.com

Received date : 9 June 2022
Acceptance date : 11 August 2022

Abstract

The purpose of this work is to accumulate information and construct a parts-of-speech tagging strategy for Kannada-language doctor-patient communications. Viterbi decoding and HMM model are utilized to accomplish this. Training data comprises of 20973 Kannada items that have been manually labeled by experts with correct parts of speech tags. Test data comprises of 450 hand gathered and pre-processed observations of doctor-patient discussions from hospitals. In order to determine the best label for the supplied Kannada term, we deployed HMM model with 27 tags. The computation precision was 91.38%. The precision achieved is reasonable compared to existing POS taggers in Kannada. Our distinctive raw data having 450 instances of Kannada-language doctor-patient communication was created specifically for this investigation.

Keywords: Hidden Markov Model, Kannada, Machine Learning, Natural language processing, parts of speech tagging.

DOI Number: 10.14704/NQ.2022.20.12.NQ77081 NeuroQuantology2022;20(12): 998-1011

1. Introduction

Tagging parts of speech is an important activity that is utilized as one of the preprocessing task in diverse NLP(Natural Language Processing) application areas (Priyadarshi & Saha, 2020).It is beneficial because it primarily enhances the performance of NLP applications (Rajani Shree & Shambhavi, 2020). Many academics are tackling this issue using a variety of methodologies, including hybrid, solutions formed on machine learning as well as rule-based strategies. Deep learning procedures for developing POS tagging models have recently emerged. Because Kannada is a large, morphologically rich, and resource-poor language, establishing a POS tagging system is challenging. Sentiment analysis, information retrieval, syntax parsing, word sense disambiguation, machine translation, and semantic parsing essentially require POS tagging, which has become core tool over NLP domain. If POS tags are not relevant for specified terms, this indicates that accuracy of POS tagging is inadequate and following work such as semantic and syntax processing may suffer as a result.

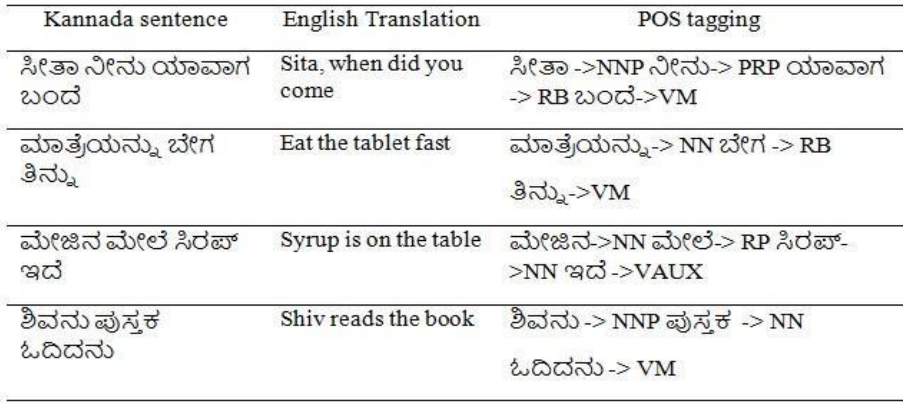
Language is a crucial communication tool. There are over 6500 languages in the globe, with 23 official languages in India. Kannada is one of them, and it is spoken mostly by people of Karnataka, which is located in India's south western area. This language is spoken natively by almost 43.7 million people. There are 49 characters in modern alphabets, with 13 vowels, 34 consonants, and two additional speech sounds. Parts of speech detection is task which identifies parts of speech for every single word in phrase. For most language processing programs, it is one of the most basic preprocessing activities. Identifying POS tags is far more difficult than it appears.

In spite of numerous strives to create suitable POS computation for the Kannada language, however development is difficult due to the language's morphological complexity. Taggers who excel at Indian languages choose a combination stochastic and machine learning approaches, as well as linguistic skills. In comparison to other languages, Kannada has a restricted corpus of linguistic material available on the internet. The Kannada language corpus must be generated manually, that could be lengthy procedure.

POS Tagging can be carried out in various ways:

1. Lexical Based Methods – Designates the POS tag to the term that appears the most frequently in the training corpus.
2. Rule-Based Methods – To generate POS tags, it utilizes rules. A rule could stipulate, for illustration, that words ending in "ed" and "ing" should be designated verbs. To enable POS tagging with regard to items that do not appear in the training sample but do appear in testing data, rule-based strategies can be used in conjunction with lexical-based methods.
3. Probabilistic Methods — This process applies tags according to the likelihood that a specific tag combination would appear. The stochastic ways for generating a POS tag are CRF(Conditional Random Fields) as well as HMM.
4. Deep Learning Methods — During POS tagging, Recurrent Neural Networks can be employed(Daniel and Martin, 2021).

Parts of speech tagging categorizes terms in corpus according to a certain part of speech relying on the word's semantics and its circumstances (Pota , Marulli, Esposito, De Pietro and Fujita, 2019). Examples of marking portions of phrase are shown in Figure 1.This research aims to collect data and devise a parts-of-speech tagging strategy for Kannada-language doctor-patient conversations.



Kannada sentence	English Translation	POS tagging
ಸೀತಾ ನೀನು ಯಾವಾಗ ಬಂದೆ	Sita, when did you come	ಸೀತಾ ->NNP ನೀನು-> PRP ಯಾವಾಗ -> RB ಬಂದೆ->VM
ಮಾತ್ರೆಯನ್ನು ಬೇಗ ತಿನ್ನು	Eat the tablet fast	ಮಾತ್ರೆಯನ್ನು-> NN ಬೇಗ -> RB ತಿನ್ನು->VM
ಮೇಜಿನ ಮೇಲೆ ಸಿರಪ್ ಇದೆ	Syrup is on the table	ಮೇಜಿನ->NN ಮೇಲೆ-> RP ಸಿರಪ್->NN ಇದೆ ->VAUX
ಶಿವನು ಪುಸ್ತಕ ಓದಿದನು	Shiv reads the book	ಶಿವನು -> NNP ಪುಸ್ತಕ -> NN ಓದಿದನು -> VM

Figure 1. Sentence POS tagging for Kannada

As a primitive phase, we physically collected 450 doctor-patient conversations from a remote medical center, which, according to the present research, has not been found anywhere else.

Manually gathering information and marking it with the appropriate tag necessitates domain expertise as well as vocabulary development. Figure 2 depicts basic schematic view.

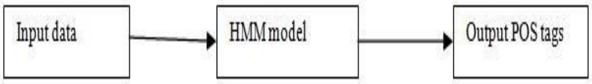


Figure 2. Basic schematic view

Objectives

- To create labeled Kannada words for training that have been manually tagged by linguistic experts.
- To collect samples of doctor patient conversation in Kannada from rural healthcare centre.
- To pre process these samples to generate test data.
- HMM model was applied to categorize each word in examples of doctor patient conversations.
- Primary goal of this study is to create data sources for Kannada in the healthcare industry

The training set comprises of 20973 Kannada items that have been cautiously and precisely labeled with the appropriate parts of speech by experts. Since Kannada is a limited resource language, we generated 450 instances of doctor-patient discussions from rural healthcare facilities to enhance the datasets in Kannada. These instances constitute as the research paper's test data. Considering 27 tags, we deployed an HMM model with viterbi interpretation. We utilized a POS label set with 21 tags that IIIT Hyderabad supplied, and we contributed 6 more manually. It is laborious to manually gather and classify the data from 450 instances of doctor-patient discussions. For our collection, we acquired a respectable precision of 91.38%. Figure 3 depicts various phases in the labeling procedure. The agglutinative character of the Kannada language enables for unrestricted word order and has a complex morphology. Kannada is a language with few assets, hence it is essential to create data sources in Kannada.

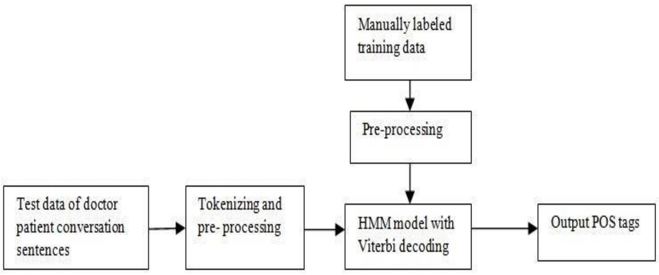

Manually labeled training data
Pre-processing
Test data of doctor patient conversation sentences
Tokenizing and pre- processing
HMM model with Viterbi decoding
Output POS tags

Figure 3. Parts of speech tagging process

2. Literature Review

They utilized CRF classifiers and wordtovec CROW model and obtained an accuracy of 82.67% and 85.88% respectively. They used 52190 words during the experiment (Priyadarshi & Saha, 2020). TDIL (Technology Development for Indian Languages) dataset with one thousand and hundred sentences was utilized and Bureau of Indian Standards tag set which owns 27 tags was also utilized. They could achieve an accuracy of 76% and 71% by using CRF++0.5 and deep learning approaches respectively (Rajani Shree & Shambhavi, 2020). Referred was a survey paper on Malayalam POS tagging (Nambiar, Leons, & Jose, 2019). Shetty and Shetty utilized HMM(Hidden Markov Model) and utilized 18000 words from online Kannada newspapers and accomplished an accuracy 95% (Shetty & Shetty, 2020). Morphological analyzer was employed and they brought about an accuracy of 92.49%. They utilized 8240 collected from internet resources (Raulji & Saini, 2019). Bidirectional Recurrent Neural networks on GENIA dataset was employed and obtained an accuracy of 94.8% (Gopalakrishnan, Soman, & Premjith, 2019). Rushali and co-authors compared machine learning and deep learning algorithms for Marathi language by exploiting 1500 sentences. They gained precision 85% for CRF, 97% precision for LSTM (Long Short Term Memory) and 85% precision for deep learning respectively (Deshmukh & Kiwelekar, 2020). Akhil and Rajimol and Anoop utilized machine learning and deep learning CUSAT Malayalam dataset and gained an accuracy of 98.78% (Akhil, Rajimol, & Anoop, 2020). They utilized BIS tag set with 36 tags for the experiment.They attempted POS tagging for the Khasi language using the Random Forest framework, and their efficiency was 92.12% (Warjri, Pakray, Lyngdoh, & Maji, 2021).

According to the experts, the 42,000 token Khasi corpora attained 0.92 of F1 score , 0.95 for test data (Warjri, Pakray, Lyngdoh, & Maji, 2021). One of the under-resourced varieties, mizo, was the target of a study. They remarked about the pos tagset that could be employed to tag the corpus in the long term (Nunsanga, Pakray, Lalngaihtuaha, & Lolit Kumar Singh, 2021). The trials were done using 5 corpora, and the authors' success rate was 91.44%, an enhancement of 3.24% over the baseline strategy (Pan & Saha, 2022). They exploited BERT to design a grammatical filtering system for the Arabic language, and their exactness was 91.69% (Saidi, Jarray, & Mansour, 2021). Table 1 depicts tagging strategies.

Table 1. Summary of various POS tagging strategies used

Method used	Dataset	Accuracy
CRF classifier and wordtovec CROW model (Priyadarshi & Saha, 2020)	52,190 words	82.67% and 85.88% respectively
CRF++0.5 and deep learning (Rajani Shree & Shambhavi, 2020)	TDIL dataset with 1100 sentences, BIS tag set - 27 tags	76% and 71% respectively
Hidden Markov Model (Nambiar et	Malayalam sentences	Survey

al., 2019)		
HMM model (Shetty & Shetty, 2020)	18000 words from online newspapers	95%
Morphological analyzer(Sanskrit) (Raulji & Saini, 2019)	8240 words from internet sources	92.49%
Bidirectional RNN,LSTM, GRU (Gopalakrishnan et al., 2019)	GENIA dataset	94.8% for bidirectional LSTM
Machine learning and deep learning algorithms (Marathi) (Deshmukh & Kiwelekar, 2020)	1500 sentences corpus from Marathi e-newspaper (32 tags)	CRF-85%,Deep learning- 85%, bi-LSTM-97%
Machine learning and deep learning techniques(Malayalam) (Akhil et al., 2020)	CUSAT Malayalam dataset (BIS tag set with 36 tags)	98.78%
Survey paper on Konkani (Rajan, Salgaonkar, & Joshi, 2020)	Survey Paper	Comparison of accuracies

Conventional Tamil taggers are imprecise and not accessible to the general public. Hence, they designed a Tamil pose tagger that, when assessed against commercially known Tamil pose taggers, had the better outcome (93.27%) (Sarveswaran & Dias, 2021). The correctness for unknown, known, and unknown terms in test sequences was 99.5%, 96.5%, and 99.8% for the experts' POS tagging for Tamil language using deep learning techniques (Visuwalingam, Sakuntharaj, & Ragel, 2021). The reliability of POS- tagging approaches for Bangla was assessed by authors. For the resource-constrained language Bangla, which has a relatively low number of annotated corpora, they applied transformation-based technique and quantitative approaches (Hasan, UzZaman, & Khan, 2007). The authors have evaluated three cutting-edge classifiers for Sinhala-POS tagging side -by- side. They validated the models using news articles corpus , official records corpus and used HMM, SVM and CRF related models (Fernando & Ranathunga, 2018). The findings revealed how to establish a POS tagger using HMM. HMM is constructed using Malayalam sentences that have already been tagged (Nambiar et al., 2019).

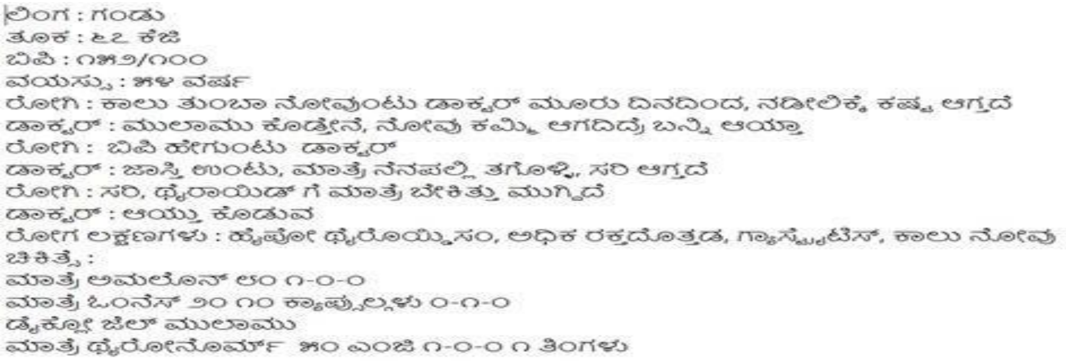
3. Methodology

We discuss dataset preparation in section 3.1 and working in section 3.2

3.1 Dataset description

We gathered Kannada-language doctor-patient dialogues from a rural allopathic clinic. We gathered 450 samples of doctor-patient talks, which included factors such as gender, age, weight, and short text messages. The severity of the illness and the duration of treatment were communicated in brief text messages. In order to gather the brief talks for this endeavor, we personally visited three physicians. Out of the collected samples obtained from 648 patients, 450 observations were utilized in this investigation. The remaining samples that weren't suitable weren't taken into account. We took into account samples pertaining to 20 common disorders, such as fever, head and cold pain, back pain, muscle pain, cough, skin conditions, stomach pain, and asthma. Figure 4 depicts the sample.

To ensure anonymity, we eliminated personal information such as names and file numbers. 20973 Kannada words were explicitly marked patient-doctor conversations. We utilized a POS tag set supplied by IIIT Hyderabad that had 21 tags and added 6 more ourselves. We and utilized in the research as training data. Doctor-patient dialogues are the test data. The Hidden Markov Model is used in conjunction with training data to get the most relevant tag for the input data. Table 2 Information on employed maximum of 27 tags as depicted in Table 3.



ಲಿಂಗ : ಗಂಡು
ತೂಕ : ೬೭ ಕೆಜಿ
ಬಿಪಿ : ೧೫೨/೧೦೦
ವಯಸ್ಸು : ೫೪ ವರ್ಷ
ರೋಗಿ : ಕಾಲು ತುಂಬಾ ನೋವುಂಟು ಡಾಕ್ಟರ್ ಮೂರು ದಿನದಿಂದ, ನಡೀಲಿಕ್ಕೆ ಕಷ್ಟ ಆಗ್ತದೆ
ಡಾಕ್ಟರ್ : ಮುಲಾಮು ಕೊಡ್ತೀನೆ, ನೋವು ಕಮ್ಮಿ ಆಗದಿದ್ರೆ ಬನ್ನಿ ಆಯ್ತಾ
ರೋಗಿ : ಬಿಪಿ ಹೇಗುಂಟು ಡಾಕ್ಟರ್
ಡಾಕ್ಟರ್ : ಜಾಸ್ತಿ ಉಂಟು, ಮಾತ್ರೆ ನೆನಪಲ್ಲಿ ತಗೊಳ್ಳಿ, ಸರಿ ಆಗ್ತದೆ
ರೋಗಿ : ಸರಿ, ಥೈರಾಯಿಡ್ ಗೆ ಮಾತ್ರೆ ಬೇಕಿತ್ತು ಮುಗ್ದಿದೆ
ಡಾಕ್ಟರ್ : ಆಯ್ತು ಕೊಡುವ
ರೋಗ ಲಕ್ಷಣಗಳು : ಹೈಪೋ ಥೈರೊಯ್ಡಿಸಂ, ಅಧಿಕ ರಕ್ತದೊತ್ತಡ, ಗ್ಯಾಸ್ಟ್ರೈಟಿಸ್, ಕಾಲು ನೋವು
ಚಿಕಿತ್ಸೆ :
ಮಾತ್ರೆ ಅಮಲೊನ್ ೮೦ ೧-೦-೦
ಮಾತ್ರೆ ಓಂನೆಸ್ ೨೦ ೧೦ ಕ್ಯಾಪ್ಸುಲ್ಗಳು ೦-೧-೦
ಡೈಕ್ಲೋ ಜೆಲ್ ಮುಲಾಮು
ಮಾತ್ರೆ ಥೈರೋನೊರ್ಮ್ ೫೦ ಎಂಜಿ ೧-೦-೦ ೧ ತಿಂಗಳು

Figure 4. Input sample data

Table 2 Information on patient-doctor conversations (test data)

Details	Count
Number of sentences in the corpus	2345
Number of words in the corpus	14854
Number of unique words in the corpus	5675

Table 3. POS tags employed

TAG(Kannada word)	**MEANING**	**EXAMPLE(English word)**
NN(Nama pada)	Common Noun	pustaka(book)
NST(Anvarthanama)	Noun Location	mele(up)
NNP(Ankita nama)	Proper Noun	dhyan(Dhyan)
PRP(Sarvanama)	Pro Noun	avalu(she)
VM(Mukyakriyapada)	Main Verb	baredanu(wrote)
VAUX(Sahayakakriyapada)	Auxiliary Verb	bareyutta(writing)
RB(Kriya visheshana)	Adverb	vegavaagi(with speed)
PSP	Post Position	jothe (together)
RP	Particles	kuda (even)
CC(Sambandarthaka)	Conjunction	mattu (and)
QF(Pramaanavachaka)	Quantifiers	bahala (very much)
WQ(Prashnartaka pada)	Question Words	yaru (who)
JJ(Nama visheshana)	Adjective	sundaravada (beautiful)

DEM(Nirnayaka pada)	Demonstrative	intaha (this kind)
QC(Sankhya vachaka)	Cardinals	ondu (one)
QO	Ordinal	ondaneya (first one)
INJ (Bhava suchaka)	Interjection	ayyo (Alas,ayyo)
CL	Group	mandi (group)
COMP(Prakaaravachaka)	Complimentizer	matte (then)
INTF(Ashcharya suchaka)	Intensifier	tumba(much)
NEG(Runathmakapada)	Negative Words	baruvudilla(wont come)
SYM(Chinhe)	Symbol	/ :
RDP(Dvirukthi)	Reduplication	begabega(fastfast)
UT(Uddharana)	Quotative	yendu (that)
XC	Compounds	agniparvatha(Mountain of Fire)
ECH(Jodu pada)	Echo Words	kadumedu(forest etc)
UNK(Gottillada pada)	Unknown Words	Hello

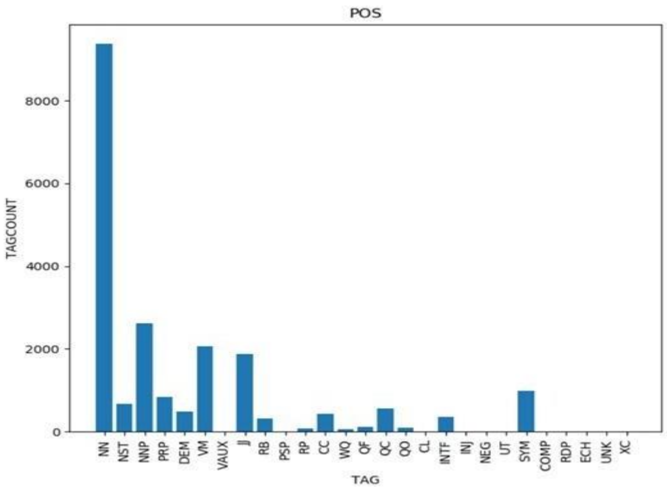


Figure 5. Tags utilized in training file

The Figure 5 shows the different tags used in the training data. The graph depicts the tags that appear the most frequently in our dataset. Only a few rare words are displayed depending on the training set's likelihood of occurrence. The Noun tag appears the most and has the greatest count of all the tags.

Working of HMM POS tagger

The training data was pre-processed, and we utilized 20973 Kannada words that were manually labeled. Words that were deemed inappropriate were omitted from the input. In the training file, the word types and number of occurrences of each tag were discovered. Figure 6 shows preprocessing and POS tagging steps. These details are utilized to compute the emission and transition matrices in the hidden markov model. Physically gathering 450 instances of Kannada doctor-patient dialogues is part of the testing procedure. UTF-8 is used to read the Kannada language input sample one line at a time (Unicode Transformation Format). Based on the training data, a Hidden Markov Model was used to mark the test sample. We used

27 user-defined tags to find the most appropriate tag.

Markov chain with a hidden markov observed and concealed states are used in the model. The observable states in our case are the Kannada words in the input, while the concealed states are the POS tags for each word. We utilize Viterbi technique to decode the hidden states given the observed states and HMM model. The rows of the emission matrix are 27 tags, while the columns are unique words retrieved from the training dataset. The transition matrix has 27 POS tags in the rows and the same tags in the columns. During deployment, we use bigrams, which means that the current word is tagged relying on the previous word.

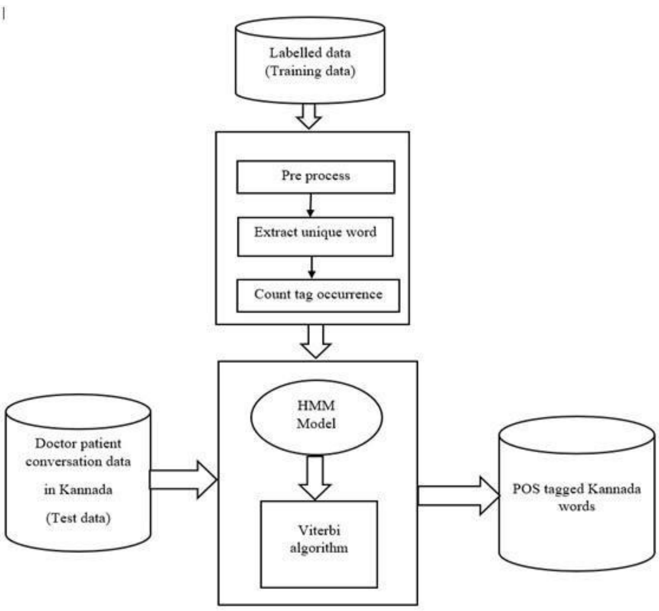


Figure 6 Preprocessing and POS tagging block diagram

During the preliminary processing of the implementation training file, undesirable words are eliminated from the file. Emission and transition matrix are initialized and upgraded, and tags are defined. The word type count is divided by the tag count for each word to replenish the emission matrix. Tag incidence is calculated by to update the transition matrix, divide the number of tags utilized in input in the test dataset. The UTF-8 encoding format was utilized to read the test file. Based on the probabilities computed in the Viterbi matrix, we read one line of test sample and apply the appropriate tag for each word in the input. For each input word, the tag with the maximum probability is adopted as the tag. The probability of a specific state in a first-order Markov chain is solely determined by the previous state. Assumption of Markov is shown in formula 1.

$$P(a_i|a_1 \ldots a_{i-1}) = P(a_i|a_{i-1}) \quad (1)$$

Second, the probability of an output data oi is solely determined by the state that generated the observation ai, not by any other events or assumptions. Independence of output is shown in formula 2.

$$P(o_i \mid a_1 \ldots a_i, \ldots, a_T, o_1 \ldots o_i, \ldots, o_T) = P(o_i|a_i) \quad (2)$$

After viewing first t samples and navigating through most likely sequence of states m1,...,mt1, The likelihood that the HMM is in state j is indicated by ht(j). Each cell's ht(j) value is determined through iteratively pursuing most likely route to that cell. As indicated in formula 3, each cell reflects the likelihood.

$h_t(j) = \max P(m_1 \ldots m_{t-1}, o_1, o_2, \ldots o_t, m_t = j \mid \lambda)$ (3)

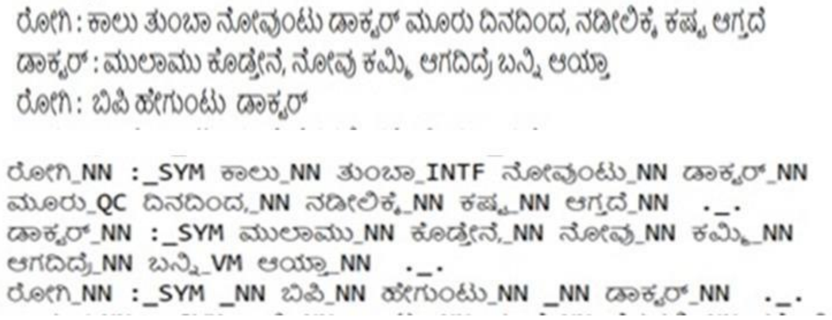
The most frequent path is depicted by taking the maximum of all preceding state sequences max m1,...,mt1. Viterbi, iteratively fills every cell. Given that we estimated possibility of being in each condition at time t-1, we estimate the Viterbi likelihood by finding most probable extensions of the routes which lead to the current cell. The value $h_j(j)$ is determined for a given state $m_j$at time t informula 4.

$h_t(j) = \max h_{t-1}(i)\, r_{ij} s_j(o_t)$ (4)

4.Results and discussion

The input data is a Kannada language doctor-patient interaction sample and the output is POS tags for each input text as depicted in Figure 7 and Figure 8. We supplied Kannada words, medical terminology, numerals, symbols and special characters as input.

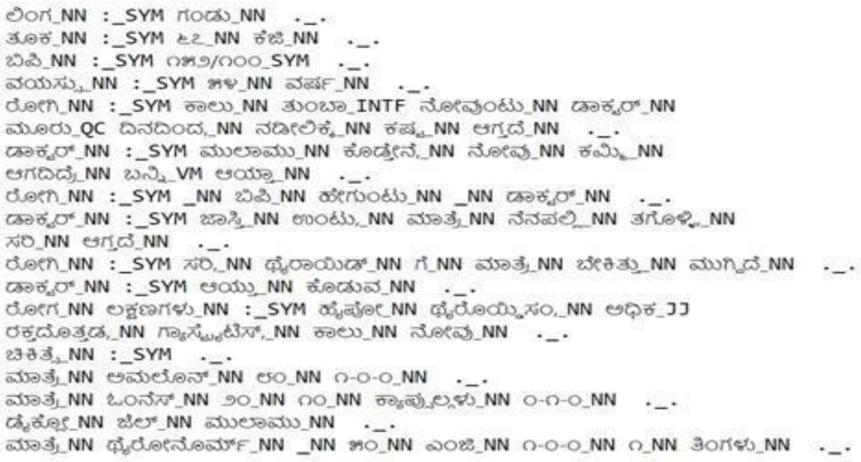
The input conversation is marked with the most relevant POS tag based on the training data. In the given input data, the Hidden Markov Model detects noun, intensifier, adjective, symbol and cardinal tags. The input data is a Kannada language doctor-patient interaction sample, and the output is POS tags for each input text. We supplied Kannada words, medical terminology, numerals, symbols and special characters as input. Figure 9 highlights count of terms that were precisely labeled in test data. Figure 10 depicts terms that were incorrectly labeled in test data. 1279 terms were wrongly labeled as depicted in Table 4.



ರೋಗಿ : ಕಾಲು ತುಂಬಾ ನೋವ್ವುಂಟು ಡಾಕ್ಟರ್ ಮೂರು ದಿನದಿಂದ, ನಡೀಲಿಕ್ಕೆ ಕಷ್ಟ ಆಗ್ತದೆ
ಡಾಕ್ಟರ್ : ಮುಲಾಮು ಕೊಡ್ತೇನೆ, ನೋವು ಕಮ್ಮಿ ಆಗದಿದ್ರೆ ಬನ್ನಿ ಆಯ್ತಾ
ರೋಗಿ : ಬಿಪಿ ಹೇಗುಂಟು ಡಾಕ್ಟರ್

```
ರೋಗಿ_NN  :_SYM  ಕಾಲು_NN  ತುಂಬಾ_INTF  ನೋವ್ವುಂಟು_NN  ಡಾಕ್ಟರ್_NN
ಮೂರು_QC  ದಿನದಿಂದ,_NN  ನಡೀಲಿಕ್ಕೆ_NN  ಕಷ್ಟ_NN  ಆಗ್ತದೆ_NN   ._.
ಡಾಕ್ಟರ್_NN  :_SYM  ಮುಲಾಮು_NN  ಕೊಡ್ತೇನೆ,_NN  ನೋವು_NN  ಕಮ್ಮಿ_NN
ಆಗದಿದ್ರೆ_NN  ಬನ್ನಿ_VM  ಆಯ್ತಾ_NN   ._.
ರೋಗಿ_NN  :_SYM  _NN  ಬಿಪಿ_NN  ಹೇಗುಂಟು_NN  _NN  ಡಾಕ್ಟರ್_NN   ._.
```

Figure 7. Input and output sample



```
ಲಿಂಗ_NN  :_SYM  ಗಂಡು_NN   ._.
ತೂಕ_NN  :_SYM  ೭೭_NN  ಕೆಜಿ_NN   ._.
ಬಿಪಿ_NN  :_SYM  ೧೫೨/೧೦೦_SYM   ._.
ವಯಸ್ಸು_NN  :_SYM  ೫೪_NN  ವರ್ಷ_NN   ._.
ರೋಗಿ_NN  :_SYM  ಕಾಲು_NN  ತುಂಬಾ_INTF  ನೋವ್ವುಂಟು_NN  ಡಾಕ್ಟರ್_NN
ಮೂರು_QC  ದಿನದಿಂದ,_NN  ನಡೀಲಿಕ್ಕೆ_NN  ಕಷ್ಟ_NN  ಆಗ್ತದೆ_NN   ._.
ಡಾಕ್ಟರ್_NN  :_SYM  ಮುಲಾಮು_NN  ಕೊಡ್ತೇನೆ,_NN  ನೋವು_NN  ಕಮ್ಮಿ_NN
ಆಗದಿದ್ರೆ_NN  ಬನ್ನಿ_VM  ಆಯ್ತಾ_NN   ._.
ರೋಗಿ_NN  :_SYM  _NN  ಬಿಪಿ_NN  ಹೇಗುಂಟು_NN  _NN  ಡಾಕ್ಟರ್_NN   ._.
ಡಾಕ್ಟರ್_NN  :_SYM  ಜಾಸ್ತಿ_NN  ಉಂಟು,_NN  ಮಾತ್ರೆ_NN  ನೆನಪಲ್ಲಿ_NN  ತಗೊಳ್ಳಿ,_NN
ಸರಿ_NN  ಆಗ್ತದೆ_NN   ._.
ರೋಗಿ_NN  :_SYM  ಸರಿ,_NN  ಥೈರಾಯಿಡ್_NN  ಗೆ_NN  ಮಾತ್ರೆ_NN  ಬೇಕಿತ್ತು_NN  ಮುಗ್ದಿದೆ_NN   ._.
ಡಾಕ್ಟರ್_NN  :_SYM  ಆಯ್ತು_NN  ಕೊಡುವ_NN   ._.
ರೋಗ_NN  ಲಕ್ಷಣಗಳು_NN  :_SYM  ಹೈಪೋ_NN  ಥೈರೊಯ್ಡಿಸಂ,_NN  ಅಧಿಕ_JJ
ರಕ್ತದೊತ್ತಡ,_NN  ಗ್ಯಾಸ್ಟ್ರೈಟಿಸ್,_NN  ಕಾಲು_NN  ನೋವು_NN   ._.
ಚಿಕಿತ್ಸೆ_NN  :_SYM   ._.
ಮಾತ್ರೆ_NN  ಅಮಲೊನ್_NN  ೮೦_NN  ೧-೦-೦_NN   ._.
ಮಾತ್ರೆ_NN  ಓಂನೆಸ್_NN  ೨೦_NN  ೧೦_NN  ಕ್ಯಾಪ್ಸುಲ್ಗಳು_NN  ೦-೧-೦_NN   ._.
ಡೈಕ್ಲೋ_NN  ಜೆಲ್_NN  ಮುಲಾಮು_NN   ._.
ಮಾತ್ರೆ_NN  ಥೈರೋನೊರ್ಮ್_NN  _NN  ೫೦_NN  ಎಂಜಿ_NN  ೧-೦-೦_NN  ೧_NN  ತಿಂಗಳು_NN   ._.
```

Figure 8. Output POS tags

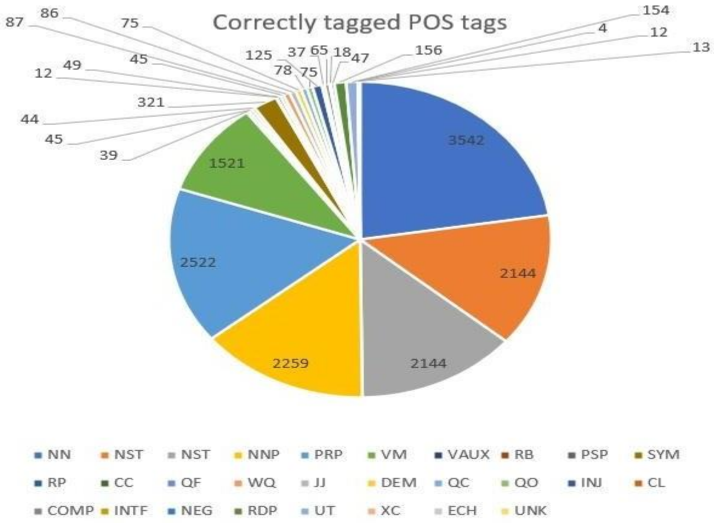


Figure 9. Terms that were precisely labeled in test data

Table 4. Incorrectly tagged words count

Incorrectly tagged words count			
NN	150 (12%)	CC	48 (4%)
NNP	138 (11%)	QF	37 (3%)
NSP	143 (12%)	DM	76 (6%)
PRP	54 (4%)	QC	45 (6%)
NEG	98 (3%)	QO	48 (4%)
QP	159 (1%)	INJ	72 (4%)
PSP	53 (2%)	JJ	19 (4%)
RB	17 (1%)	VAUX	14 (6%)
VM	96 (13%)	OTHERS	12 (1%)

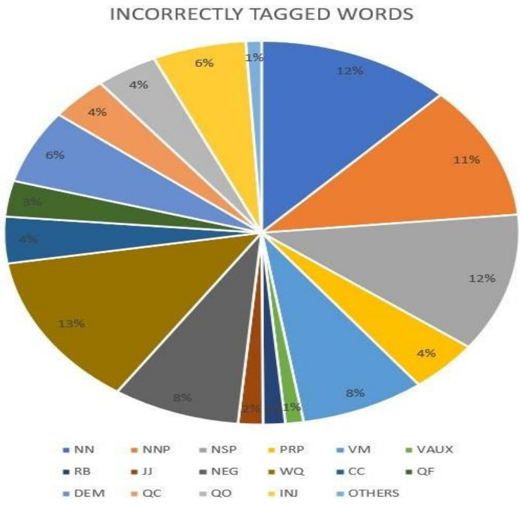


Figure 10. Percentage of terms that were incorrectly labeled in test data

The correctness of POS Tagger is evaluated by comparing the system-tagged words to the hand-annotated words and estimating the consistent POS tag assignments as depicted in Equation (5).

$$\text{Accuracy} = \frac{\mathit{Total\ number\ \ of\ words\ appropriately\ \ tagged}}{\text{Total number of words}} \quad (5)$$

With a training dataset of 20,973 words, the bigram language model based POS tagger reported consistent precision of 91.38%. Comparison of precision with suggested model is depicted in Table 5.Utilizing logic to coding problems in native language (Rao, GB, Guntnur, Reddy, & KR, 2020). The researchers' strategies and techniques for interpreting Indian minority language get investigated (Harish & Rangan, 2020). Deep learning accounted for 68% of suggested Classifiers solutions, although machine learning and hybrids accounted for remaining 20% and 12% of designs (Chiche & Yitagesu, 2022). With the manually compiled medical dataset, we were able to get a good precision of 91.38% for Kannada language. Increasing the dataset would improve the accuracy of the current system.

Table 5. Comparison of accuracies with proposed model.

Method used	Accuracy
CRF classifier and wordtovec CROW model (Priyadarshi & Saha, 2020)	82.67% and 85.88% respectively
CRF++0.5 and deep learning (Rajani Shree & Shambhavi, 2020)	76% and 71% respectively
HMM model and existing dataset (Shetty & Shetty, 2020)	95%
Bidirectional RNN,LSTM, GRU (Gopalakrishnan et al., 2019)	94.8% for bidirectional LSTM
Machine learning and deep learning algorithms (Marathi) (Deshmukh & Kiwelekar, 2020)	CRF-85%,Deep learning-85%, bi-LSTM-97%
Machine learning and deep learning techniques(Malayalam) (Akhil et al., 2020)	98.78%
Parts of speech of Odia language using deep learning (Dalai, Mishra, & Sa, 2022)	94.58%
Proposed HMM model with manually created medical dataset (Kannada)	91.38%

5. Conclusion and overlook

We depicted literature on POS tagging for Kannada as well as other Indian languages, and HMM model has been deployed for Kannada POS tagging. A collection of 20973 properly labeled Kannada words has been developed by our expertise which is training data. We acquired 450 instances of doctor-patient interactions for testing. To pick the best label for the input term, we utilized HMM with 27 labels. Repository of 20973 appropriately labeled Kannada terms has been generated by discussion with linguistic experts. To locate the best label for input word, we utilized HMM with 27 descriptors. The experiment yielded 91.38% precision by utilizing manually generated dataset. Kannada is dialect with constrained resources, hence our suggested effort presents the novel notion of building datasets for Kannada in the healthcare sector.By employing deep learning algorithms and expanding the dataset, precision can be boosted.

References

Akhil, K. K., Rajimol, R., & Anoop, V. S. (2020). Parts-of-Speech tagging for Malayalam using deep learning techniques. International Journal of Information Technology, 12(3), 741-748. https://doi.org/10.1007/s41870-020-00491-z

Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. Journal of Big Data, 9(1), 1-25. https://doi.org/10.1186/s40537-022-00561-y

Dalai, T., Mishra, T. K., & Sa, P. K. (2022). Part-of-Speech Tagging of Odia Language Using statistical and Deep Learning-Based Approaches. arXiv preprint arXiv:2207.03256. https://doi.org/10.48550/arXiv.2207.03256

Deshmukh, R. D., & Kiwelekar, A. (2020, March). Deep learning techniques for part of speech tagging by natural language processing. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 76-81). IEEE. https://doi.org/10.1109/ICIMIA48430.2020.9074941

Fernando, S., & Ranathunga, S. (2018, May). Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon) (pp. 96-101). IEEE. 10.1109/MERCon.2018.8421997

Gopalakrishnan, A., Soman, K. P., & Premjith, B. (2019, July). Part-of-speech tagger for biomedical domain using deep neural network architecture. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE. https://doi.org/10.1109/ICCCNT45670.2019.8944559

Harish, B. S., & Rangan, R. K. (2020). A comprehensive survey on Indian regional language processing. SN Applied Sciences, 2(7), 1-16. https://doi.org/10.1007/s42452-020-2983-x

Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla. In Advances and innovations in systems, computing sciences and software engineering (pp. 121-126). Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6264-3_23

Nambiar, S. K., Leons, A., & Jose, S. (2019, December). Natural Language Processing Based Part of Speech Tagger using Hidden Markov Model. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp.

782-785). IEEE. 10.1109/I-SMAC47947.2019.9032593

Nunsanga, M. V., Pakray, P., Lalngaihtuaha, M., & Lolit Kumar Singh, L. (2021). Part-of-speech tagging in Mizo language: A preliminary study. In Data Intelligence and Cognitive Informatics (pp. 625-635). Springer, Singapore. 10.1007/978-981-15-8530-2_49

Pan, S., & Saha, D. (2022). Performance Evaluation of Part-of-Speech Tagging for Bengali Text. Journal of The Institution of Engineers (India): Series B, 103(2), 577-589. https://doi.org/10.1007/s40031-021-00630-5

Priyadarshi, A., & Saha, S. K. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. Computer Speech & Language, 62, 101054. https://doi.org/10.1016/j.csl.2019.101054

Rajani Shree, M., & Shambhavi, B. R. (2020). POS tagger model for Kannada text with CRF++ and deep learning approaches. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), 485-493. https://doi.org/10.1080/09720529.2020.1728902

Rajan, A., Salgaonkar, A., & Joshi, R. (2020). A survey of Konkani NLP resources. Computer Science Review, 38, 100299. https://doi.org/10.1016/j.cosrev.2020.100299

Rao, V., GB, S., Guntnur, S., Reddy, S., & KR, P. (2020, November). Using Natural Language Processing to Translate Plain Text into Pythonic Syntax in Kannada. In Proceedings of the Future Technologies Conference (pp. 664-680). Springer, Cham. DOI: 10.1007/978-3-030-63128-4_51

Raulji, J. K., & Saini, J. R. (2019). Sanskrit lemmatizer for improvisation of morphological analyzer. Journal of Statistics and Management Systems, 22(4), 613-625. https://doi.org/10.1080/09720510.2019.1609186

Saidi, R., Jarray, F., & Mansour, M. (2021, June). A BERT based approach for Arabic POS tagging. In International Work-Conference on Artificial Neural Networks (pp. 311-321). Springer, Cham. 10.1007/978-3-030-85030-2_26

Sarveswaran, K., & Dias, G. (2021, December). Building a Part of Speech tagger for the Tamil Language. In 2021 International Conference on Asian Language Processing (IALP) (pp. 286-291). IEEE. 10.1109/IALP54817.2021.9675195

Shetty, S., & Shetty, S. (2020). Text pre-processing and parts of speech tagging for Kannada language. Journal of Xi'an University of Architecture & Technology, 12(II), 1286-1291. http://www.xajzkjdx.cn/gallery/120-feb2020.pdf

Visuwalingam, H., Sakuntharaj, R., & Ragel, R. G. (2021). Part of Speech Tagging for Tamil Language Using Deep Learning. In 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS) (pp. 157-161). IEEE. 10.1109/ICIIS53135.2021.9660738

Warjri, S., Pakray, P., Lyngdoh, S., & Maji, A. K. (2021). Adopting conditional random field (crf) for khasi part-of-speech tagging (kpost). In Proceedings of the International Conference on Computing and Communication Systems (pp. 75-84). Springer, Singapore. 10.1007/978-981-33-4084-8_8

Warjri, S., Pakray, P., Lyngdoh, S. A., & Maji, A. K. (2021). Part-of-speech (pos) tagging using conditional random field (crf) model for khasi corpora. International Journal of Speech Technology, 24(4), 853-864. https://doi.org/10.1007/s10772-021-09860-w