Kannada Parts of Speech Tagging for medical dataset

Sudiksha Santosh Nayak[0000−0002−9904−1720], Swasthika[0000−0003−1470−6605], Swathi Mithanthaya[0000−0003−0535−5524], and Swathi S Rao[0000−0001−9256−9277]

Dept. of Computer Science and Engineering, NMAM Institute of Technology, Karkala, Karnataka, India
sudiksha132@gmail.com, swasthikadevadiga2@gmail.com, swathi.mithanthaya@gmail.com swathirao029@gmail.com

Abstract. Parts of Speech (POS) tagging is one of the basic text processing tasks of Natural Language Processing (NLP). It is the process of assigning the part of speech tag to each and every word in a sentence. In this paper, we have presented POS tagger for South Indian language Kannada, using Hidden Markov Model. POS tagger has been developed using Python Programming Language. The result is based on experiment conducted on a large number of data sets where the training data consisted of 20,000 kannada words with their respective POS tags and around 240 testing data sets consisted of different conversations between a doctor and patient. The tag set consists of 27 different tags which are used to assign the POS.

Keywords: POS Tagging · Hidden Markov Model · Viterbi

1 Introduction

Language is a vital tool for communication. There are around 6500 languages in the world and in India we have 23 official languages. One of them is Kannada which is spoken predominantly by the people of Karnataka which is in the south western region of India. Over 43.7 million people are the native speakers of this language. The modern alphabets have 49 characters which comprise of 13 vowels, 34 consonants and 2 other speech sounds. Parts of Speech tagger is used to assign POS tag to each word in a sentence. Identifying POS tags is much more complicated than simply mapping words to their part of speech tags. This is because POS tagging is not something that is generic.
A brief survey about various related works is in section 2, followed by section 3 consisting of various challenges concerned with Hidden Markov Model. Section 4 and 5 explains the datasets used and implementation of POS tagging using HMM model. In the subsequent section, we find the result of our experiment. The paper concludes with future possible enhancements.

2 Related Work

In paper,Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition[2], the authors provide Introduction to NLP,computational linguistics.They have connected traditional foundations,recent developments and trends and stacked it all together.

In paper,The effects of part-of-speech tagsets on tagger performance [3], the author modifies the tagset to form a new finer-grained tagset, which more accurately reflect the basal linguistic distinctions which helps to more precisely determine the identity of surrounding tags.

In paper,Comparison of different POS Tagging Techniques for Bangla[1], the author uses annotated tagging set corpus for pos tagging of Bangla language data usig Brill's supervised tagger method. They have used n-gram HMM model to train the dataset.

In paper,Part-of-Speech Tagging with Minimal Lexicalization[6], the authors prospect the effect of eliminating redundancy and thoroughly reduce the size of feature vocabularies.Finding a small but linguistically motivated set of suffixes results in improved cross-corpora generalization.

In paper,A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing, [7], the authors provide a brief description about various attempts in POS taggers in Indian languages and different Tagset developed for Indian languages.

In paper,Maximum Entropy Approach to Kannada Part Of Speech Tagging [5] , the authors define a probabilistic classifier technique of Maximum Entropy model and is evaluated for the tagging of Kannada sentences which is agglutinate, morphologically very rich but resource poor.

In paper,Named Entity Recognition in Biomedical Texts using an HMM Model[8], the author work shows that the word similarity-based smoothing can improve the performance by using huge unlabeled data they focus on the fact that that word similarity is a potential method to automatically get word formation.

In paper, A Systematic Review of Hidden Markov Models and Their Applications[4]. the author shows the significant trends in the research on hidden Markov model variants and their applications they evaluate hidden Markov model variants in work in various application fields.

3 Challenges

The major challenges concerned with Hidden Markov Model are categorised into three definite sects

1.Evaluation problem

Finding the probability of the sequence which simply refers to finding likelihood of the data.Given the observation sequence V=v_1,v_1,v_1,...,v_t and a model $\theta = (A, B, \pi)$, *how to we efficiently compute* $P(V|\theta)$, *the probability of observation of sequence, given the model*?

2.Decoding

Finding the most likely sequence of hidden state, given an observation sequence which is we take parts of speech as the hidden state and the actual word in the sentence is observation. Given the observation sequence V=v_1,v_1,v_1,...,v_t and the model θ, *how do we choose a corresponding state sequence*S =s_1,s_1,s_1,...,s_t which is optimal (i.e., best "explains" the observations)?

3.Learning

Maximizing the likelihood of sequence with respect to parameters which is how do we adjust the model parameters $\theta = (A, B, \pi)$ *to maximize* $P(V|\theta)$?

Some of the challenges faced regarding POS tagging are:

Presence of lexical ambiguous words in the corpus which means presence of two or more possible meanings for a single word which leads to a serious issue while tagging. Normalization of the corpus which is precise technique of transforming text a single canonical form must be undertaken. Concerns regarding compound words and multi word expression. Presence of abbreviation in the corpus makes the tagging complex.

Presence of special symbol or characters which are not in the root language of corpus,hence its very necessary to specially deals with such words for the better accuracy of the model. Ambiguity in attachment of the preposition in corpus, leading to the different interpretation of a same word.

When unknown words which is out of context of tag set are seen, the approach can be to assign suffix and calculate the probability that the suffixed word with a particular tag occurs in a sequence. To deal with sparse data where the probability is frequency of words is zero smoothing techniques should be incorporated.

4 Dataset

These are doctor diagnosis chatbot chat details, between the patient and the chatbot/system/doctor (in this case). These are the preliminary questions asked once a communication has been established by the user and the system. These questions are basic information about the users, some of which include information such as gender, age, the kind of diagnosis the patient has and the system (doctor) asking related question to the same, other information collected contains weight, blood pressure, temperature check etc, depending on the patients diagnosis.

A simple example of the dataset is (note the data set is in Kannada and only for explanation purpose English words have been used)

Age: 52
Gender: Female
Doctor: Come sit
Patient: Hello Doctor
Doctor: Hello
Patient:My ankle hurts
Doctor:oh
Doctor:Do you have BP

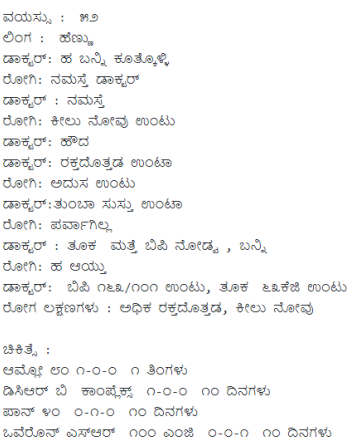

Fig. 1. Actual Input without any tags in Kannada

Patient:yes i do
Doctor:Do you get exhausted early ?
Patient:No
Doctor:Lets weigh you and check your BP
Patient:Ok
Doctor:BP - 163/101 and you weigh - 63kg
Patient Symptoms: High BP and ankle pain

prescription:
Aamlo 80 1-0-1 1 month
DCR B Complex 1-0-0 10 days
Paan 40 0-1-0 10 days
Overon SR 100mg 0-0-1 10 days

What our model predicted :
age VM : SC NN $52_N N$
$genderVM : SCNNfemaleNN$
$Doctor : NNYesNNcomeVMsitFS$
$Patient : NNNamasteNNDoctor_N N$
$DoctorVM : SCNamasteNN$
$Patient : NNankleNNpainVMhaveVM$
$Doctor : NNyes?NN$
$Doctor : NNBPNNhaveNN$
$Patient : NNyesNNhaveVM$
$Doctor : alotNNtirednessNNhaveNN$
$Patient : NNalrightNN$

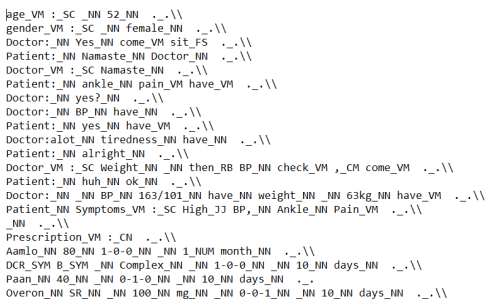

Fig. 2. Output of a data with POS tagged

$DoctorVM : SCWeightNNNNthenRBBPNNcheckVM, CMcomeVM$
$Patient : NNhuhNNokNN$
$Doctor : NN_NNBP_NN163/101_NNhave_NNweight_NN_NN63kg_NNhave_VM$
$PatientNNSymptomsVM : SCHighJJBP, NNAnkleNNPainVM$

$PrescriptionVM : CN$
$AamloNN80NN1-0-0NNNN1NUMmonthNN$
$DCRSYMBSYMNNComplexNNNN1-0-0_NNNN10NNdaysNN$
$PaanNN40NNNN0-1-0NNNN10NNdaysNN$
$OveronNNSRNNNN100NNmgNNNN0-0-1NNNN10NNdaysNN$

Note - The sentences differ a little as the output is the direct translation of the sentences from Kannada.

5 Implementation

Approach for POS tagging:

The technique used here for Kannada POS tagging is supervised and stochastic hidden markov model (HMM) using the algorithm known as Viterbi. We have used the stochastic approach by calculating the frequency of occurrence of tags for a particular word in the training dataset. Furthermore, the calculation of probabilities of a given sequence of tags. It determines the best tag by comparing to the previous n-tags. This method is called n-gram and it is implemented using Viterbi algorithm.

The model:

HMM is called Hidden because it uses word sequences and exact sequence of tags that generated this word sequence is unknown. It is called Markov as

it is based on the Markovian assumption that the current tag depends only on the previous tags. The HMM model trains on tagged dataset to find out the transition and emission probabilities.

For a sequence of words w, HMM determines the sequence of tags t using the formula: t = argmax P(w, t). The computation of this formula is very expensive as all possible tag sequences are required to be checked in order to find the sequence that maximizes the probability. So, a dynamic programming approach known as the Viterbi Algorithm is used to find the optimal tag sequence. Viterbi is a search algorithm makes use of m- Maximum Likelihood Estimates (MLE) where m represents the number of tags of a particular word. We find Viterbi path that provides the word sequences without using the complex HMM formula.

The algorithm takes 2 matrices as input: Emission is a matrix storing the log-probability of observing word and tag. Transition is a matrix storing the transition log-probability from the previous tag to the current tag.

The implementation:

We have used VSCode to implement the following
1) Open the training set and read the file with encoding 'utf-8' using codecs library.
2) Pre-process the file by removing all unnecessary tags and words.
3)Initialise the tags list with all tags used in the training set.
4)Unique words are extracted from the training data by calculating the count of each tag and word type in the training set.
5) Count of occurrence of each tag is calculated.
6) Emission & Transmission matrix are initialized and computed.
7) Testing data is read from the terminal and the file is opened with encoding 'utf-8' using codecs library.
8) Viterbi decoding is computed.
9) Tagged text is printed on to output file.

6 Results

Since, testing and training data are stored in different files, we will be storing every result in a separate file in a folder called Output. Consider that the code is stored in a file called code.py. We have used the library sys to read the testing file from the terminal. We will be passing test data as a parameter for the code file. python code.py testdata.txt The format of output data will be that every Kannada word will be followed by an underscore and then the POS tag. See Fig.1.for the sample output.

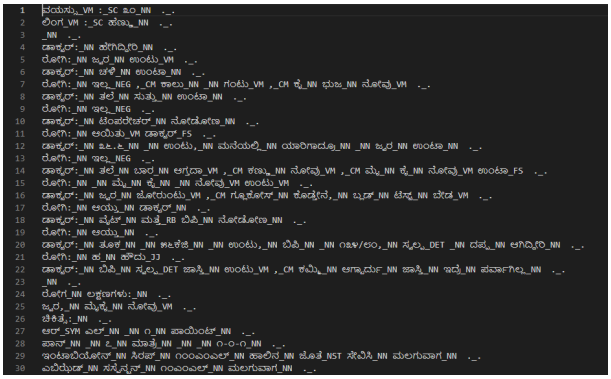

Fig. 3. Output of a test data with POS tagged

7 Conclusions and future work

Natural Language is a way of communication. One of the most important activities in processing natural languages is Part of Speech tagging. In this POS Tagging project we have assigned a Part of Speech tag to each word in a paragraph. This project mainly focuses on POS tagging for Kannada paragraphs using the HMM(Hidden Markov Model). A simple task like Part of Speech Tagging can easily be done by Hidden Markov Model. Viterbi Algorithm helps in effectively search for the best hidden sequence given an observed sequence.

POS tagging is one of the simplest, statistical model for many NLP application. POS Tagging is in its initial stage of linguistics, text analysis like information retrieval, machine translator, text to speech synthesis, information extraction etc. Many companies like Google and Microsoft are concentrating on Natural language processing applications and have developed many applications for the same. The POS tagger described here is very simple and efficient for automatic tagging. The necessity of a linguistic background and manually constructing the rules are the main drawbacks of the rule based systems. The problem with this approach is that it can come up with sequences of tags for sentences that are not acceptable according to the grammar rules of a language. The performance of the current system is good. We believe that future enhancements of this work would be to improve the tagging accuracy and implementing on more languages.

References

1. Hasan, F.M., UzZaman, N., Khan, M.: Comparison of different pos tagging techniques (n-gram, hmm and brill's tagger) for bangla. In: Elleithy, K. (ed.) Advances and Innovations in Systems, Computing Sciences and Software Engineering. pp. 121–126. Springer Netherlands, Dordrecht (2007)

2. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, USA, 1st edn. (2000)
3. MacKinlay, A.D.: The effects of part-of-speech tagsets on tagger performance (2005)
4. Mor, B., Garhwal, S., Kumar, A.: A systematic review of hidden markov models and their applications. Archives of Computational Methods in Engineering **28** (05 2020). https://doi.org/10.1007/s11831-020-09422-4
5. R, S., P, R.K., G, R.: Article: A maximum entropy approach to kannada part of speech tagging. International Journal of Computer Applications **41**(13), 9–12 (March 2012), full text available
6. Savova, V., Peshkin, L.: Part-of-speech tagging with minimal lexicalization (01 2004). https://doi.org/10.1075/cilt.260.18sav
7. V.Chitraa, Thanamani, D.S.: Article: A novel technique for sessions identification in web usage mining preprocessing. International Journal of Computer Applications **34**(9) (November 2011), full text available
8. Zhao, S.: Named entity recognition in biomedical texts using an hmm model. In: NLPBA/BioNLP (2004)