STUDY OF PART OF SPEECH TAGGING

Vaditya Ramesh
111CS0116

Department of Computer Science
National Institute of Technology, Rourkela
May, 2015

STUDY OF PART OF SPEECH TAGGING

Thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Computer Science and Engineering

by

Vaditya Ramesh
111CS0116

Under the supervision of

Prof. Ramesh Kumar Mohapatra

Department of Computer Science
National Institute of Technology, Rourkela
May, 2015

Department of Computer Science and Engineering
National Institute of Technology, Rourkela
Rourkela-769008, Odisha, India

May, 2015

Certificate

This is to certify that the work in the project entitled *Study of Parts of Speech Tagging* by *Vaditya Ramesh* is a record of his work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Bachelor of Technology* in *Computer Science and Engineering*.

Date:
Place:

Prof. Ramesh Kumar Mohapatra
Dept. of Computer Science and Engineering
National Institute of Technology Rourkela - 769008

Acknowledgement

I am indebted to my guide Prof. R. K. Mohapatra for giving me an opportunity to work under his guidance. Like a true mentor, he motivated and inspired me through the entire duration of my work.

Last but not the least, I express my profound gratitude to the Almighty and my parents for their blessings and support without which this task could have never been accomplished.

Date:

Place:

Vaditya Ramesh
Dept. of Computer Science and Engineering
National Institute of Technology Rourkela - 769008

Abstract

In the area of text mining, Natural Language Processing is a rising field. So if a sentence is an unstructured to make it a suitable organized organization. Grammatical feature Tagging is one of the preprocessing steps which perform semantic examination.

Parts of Speech labeling appoints the suitable grammatical feature and the lexical classification to each word in the sentence in Natural dialect. It is one of the key undertakings of Natural Language Preparing. Parts of Speech labeling is the first venture taking after which different procedures as in chunking, parsing, named substance acknowledgment and so on. An adjustment of different machine learning strategies are connected in particular Unigram Model and Hidden Markov Model (HMM).

The huge focuses realized by thesis can be highlighted below:
•Use of Unigram and Hidden Markov Model for parts of Speech Tagging and analyzing their performance

•**Keywords:** Brown Corpus, POS Tagger, Unigram Model and Hidden Markov.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

There is a wide range and center zones in Human Language Technology (HLT). This incorporates regions, for example, Natural Language Processing (NLP), Speech Recognition, Machine Translation, Text Generation and Text Mining. A characteristic Language understanding framework must have information about what the words mean, how words consolidate to frame sentences, how the word implications join to from sentence implications. A Part-Of-Speech Tagger (POS Tagger) is Software. It peruses content information and appoints parts of discourse to every word, for example, thing, verb, modifier, pronoun, and relational word and so on. Grammatical feature marks are utilized to determine the inward structure of a sentence. The methodology of partner names with every token in content is called labeling and the marks are called labels. The gathering of labels utilized for a specific errand is known as a label set. Grammatical feature labeling is the most well-known sample of labeling. Sample ('great', 'Adj') in this case word great showing token and Adj is a Adjective of token and it demonstrates Tagging of that token.

Unigram and Hidden Markov taggers label every word (tokens) with the tag i.e. destined to run with the token's sort. It utilizes a preparation corpus to choose which tag is in all probability for every sort. Specifically, it accepts that the label that happens most habitually with a sort is the no doubt tag

for that sort. Part-of-speech tags divide words of sentence into categories, based on how they can be combined to form semantic sentences. Case in point, articles can consolidate with things, however not verbs. Grammatical form labels additionally give data about the semantic substance of a word. For instance, things normally express" things," and relational words express relationship between things. Most of the part-of-speech tag sets make use of the same basic categories, such "noun," "verb," "adjective," and "preposition", "adver". On the other hand, label sets contrast both in how finely they partition words into classifications; and in how characterize their classes. Case in point, "is" may be labeled as a verb in one label set; however as a type of "to be" in another label set. This variety in label sets is sensible, since grammatical feature labels are utilized as a part of diverse routes for distinctive errands. Bellow mentioned basic tag set of a part-of -speech.

1.1 Application of Part Of Speech (POS) tagger

Natural language processing task has to remove parts of speech ambiguity. As so, it can be considered as the first step of language understanding. Further processes may include Parsing, Morphological Analysis, and Chunking etc. Tagging is often a necessity for many applications as in Speech Analysis and Recognition, Machine translation, lexical analysis and information retrieval. The Applications of POS labeling are specified underneath:

1. Fundamentally, the objective of a grammatical form tagger is to allot etymological (generally linguistic) data to sub-sentential units. Such units are called tokens and the greater part of the times relate to words and images (e.g. accentuation). 2. It utilizes as a part of parsing and content to-discourse transformation. 3. Data extraction, on making an inquiry to the master framework, a great deal of data about the parts of discourse can be recovered. Along these lines, if one needs to hunt down records that contain "building"

as a verb, one can include extra data that evacuates the likelihood of the word to be recognized as a thing. 4. Discourse Recognition and combination, an amazing measure of data is separated about the word and its neighbors from its parts of discourse. This data will be helpful for the dialect model.

Chapter 2

Literature Review

2.1 Part Of Speech Tagger

Grammatical feature (POS) labeling is the procedure of marking a Part-of-Speech or other lexical class marker to every word in a sentence. It is like the procedure of tokenization for coding languages. Consequently POS labeling is considered as a vital process in discourse acknowledgment, characteristic dialect parsing, morphological,parsing, data recovery and machine interpretation. Distinctive methodologies have been utilized for Part-of-Speech (POS) labeling, where the eminent ones are principle based, stochastic, or change based learning approaches. Principle based taggers attempt to allot a tag to every word utilizing an arrangement of manually written rules. These guidelines could determine, for occurrence that a word taking after a determiner and a modifier must be a thing. This implies that the arrangement of principles must be appropriately composed and checked by human specialists. The stochastic (probabilistic) approach utilizes a preparation corpus to pick the most likely tag for a word. There are a couple of different strategies which utilize probabilistic methodology for POS Labeling, for example, the Tree Tagger. At last, the change based methodology joins the standard based methodology and factual methodology.

It picks the undoubtedly tag in view of a preparation corpus and after that

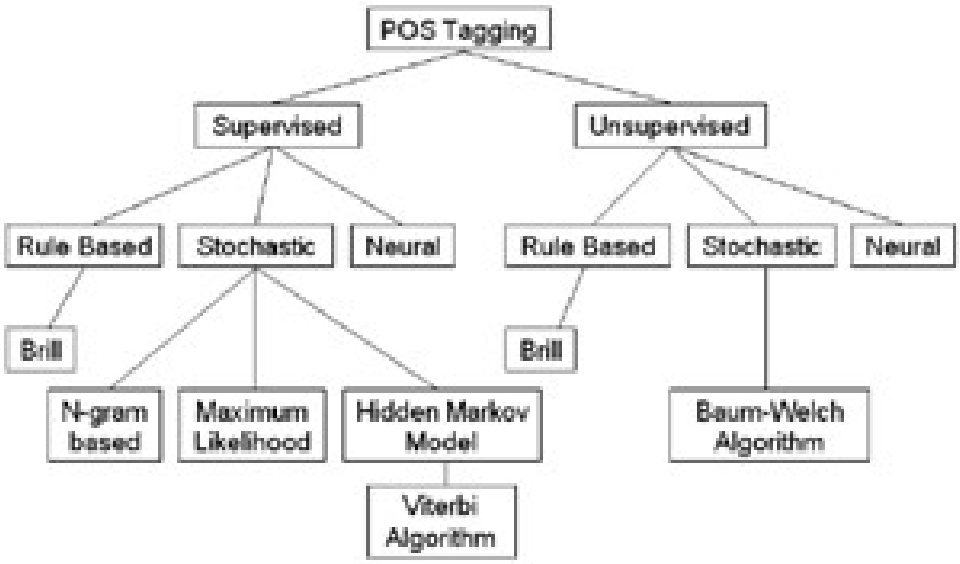
applies a certain arrangement of standards to see whether the tag should be changed to whatever else. It spares any new decides that it has learnt in the process, for future utilization. One case of a viable tagger in this class is the Brill tagger .Supervised POS Tagging, where a pretagged corpus is an essential. On the other hand, there is the unsupervised POS labeling strategy and it doesn't require any pretagged corpora. Koskenniemi additionally utilized a tenet based methodology actualized with limited state machines. Greene and Rubin have utilized a standard based approach in the TAGGIT program, which was a guide in labeling the Brown corpus. Derouault and Merialdo have utilized a bootstrap system for preparing. At to begin with, a generally little measure of content was physically labeled and used to prepare an incompletely exact model. The model was then used to tag more content, and the labels were physically redressed and afterward used to retrain the model. Church utilizes the labeled Brown corpus for preparing. These models include probabilities for every word in the vocabulary and henceforth an expansive labeled corpus is needed for a dependable estimation. Jelinek has utilized Hidden Markov Model (HMM) for preparing a content tagger.

Parameter smoothing can be helpfully attained to utilizing the technique for 'erased interjection' in which weighted evaluations are taken from second and first-arrange models and a uniform likelihood appropriation. Kupiec utilized word equality classes (alluded to here as uncertainty classes) taking into account parts of discourse, to pool information from individual words. The most normal words are still spoken to independently, as adequate information exist for strong estimation. Yahya O. Mohamed Elhadj presents the advancement of an Arabic grammatical form tagger that can be utilized for investigating and commenting conventional Arabic writings, particularly the Quran content. The created tagger utilized a methodology that joins morphological examination with Hidden Markov Models (HMMs) based-on the Arabic sentence structure. The morphological investigation is utilized to decrease the

measure of the dictionary labels by dividing Arabic words in their prefixes, stems and postfixes; this is because of the way that Arabic is a derivational dialect. Then again, HMM is used to speak to the Arabic sentence structure to consider the semantic blends. In the late writing, a few ways to deal with POS labeling taking into account factual also, machine learning systems are connected, including Hidden Markov Models, Most extreme Entropy taggers, Transformation–based learning, Memory–based learning, Decision Trees, and Support Vector Machines. The majority of the past taggers have been assessed on the English WSJ corpus, utilizing the Penn Treebank set of POS classes.

2.2 Classification of Part Of Speech Tagging

Figure 2.1: Classification of Part Of Speech Tagging



2.3 Supervised models

The Supervised POS Tagging models oblige a pre annotated Corpus which is used for planning to learn information about the tagset, word-mark frequencies, guideline sets, et cetera [1]. The execution of the models generally augments with addition in the compass of the corpus.

2.4 Unsupervised models

The unsupervised POS Tagging models don't oblige a preannotated corpus. Maybe, they use advanced computational techniques like the BaumWelch count to therefore actuate tag sets, change principles, et cetera. In perspective of this information, they either process the probabilistic information needed by the stochastic taggers or impel the legitimate rules needed by rule based systems or change based structures [1, 2]. Both the directed and unsupervised models can be further described into the going with classes.

2.5 Stochastic models

The stochastic models consolidate repeat, probability or estimations. They can be in perspective of differing procedures, for instance, n-grams, most noteworthy likelihood estimation (MLE) or Hidden Markov Models (HMM). HMM-based systems oblige appraisal of the argmax formula, which is greatly sumptuous as all possible name groupings must be checked, with a particular finished objective to find the gathering that increases the probability. So a component programming procedure known as the Viterbi Calculation is used to find the perfect name progression [3]. There have in like manner been a couple of studies utilizing unsupervised learning for setting up a HMM for POS Labeling. The most by and large known is the Baum-Welch estimation, which can be used to set up a HMM from un-illuminated data. Moreover,

eventually, both directed and unsupervised POS Labeling models can be considering neural frameworks.

Chapter 3

POS Tagging using Unigram Model

3.1 POS Tagging using Unigram Model

The Unigram Tagger class executes a straightforward factual labeling calculation: for every token, it appoints the label that is in all probability for that token's sort. For instance, it will allot the label "JJ" to any event of "incessant," since "regular" is utilized as a modifier (e.g. "an incessant word") more regularly than it is utilized as a verb. Before a Unigram Tagger can be utilized to label information, it must be prepared on a preparation corpus. It utilizes this corpus to figure out which labels are most regular for every word. Unigram Taggers are prepared utilizing the train strategy, which takes a labeled corpus. Unigram Tagger will allot the default label none to any token whose sort was not experienced in the preparation information. Before labeling information is ought to be tokenized. Unigrams: P (t) =f (t)/N In this comparison f (t) speaks to the recurrence of label t and N speaks to the aggregate number of tokens in the corpus. Brown corpus: Brown corpus is a database .It contains standard information. Chestnut corpus of standard American English was the first of the current PC clear, general corpora. It was assembled by W.N.Francis and H.kucera .The corpus comprises of one million expressions of American English writings imprinted in 1961.

Chapter 4

Hidden Markov Model

4.1 Tagging with Hidden Markov Model

In this part is portrayed the HMM based computation for Parts Of Speech labeling. Shrouded Markov Model is a champion amongst the adequately used dialect show (1-gram...n-gram) for inferring names which uses uncommon measure of information about the dialect, divided from straightforward setting related information. A HMM is a stochastic based build which could be used to handle the arrangement issues that have a state grouping structure.

The model has various interconnected states associated by their move likelihood. A move likelihood is the likelihood that framework moves starting with one state then onto the next. A procedure starts in one of the states, and moves to another state, which is represented by the move likelihood. A yield image is transmitted as the procedure moves starting with one state then onto the next. These are otherwise called the Observations. HMM fundamentally yields a succession of images. The transmitted image relies on upon the likelihood dissemination of the specific state. Anyhow, the careful succession of states concerning a normal perception arrangement is not known (covered up).

Defining of an HMM:

Success a Weighted Finite-state Automaton (WFSA)

-Each move curve is connected with likelihood

-The entirely of all bends active from a solitary hub is 1

-Markov chain is a WFSA in which an info sting remarkably focus way through the automation

-Hidden Markov Model (HMM) is a somewhat distinctive case in light of the fact that some data (past POS lables) is obscure (or covered up)

-HMM comprises of the accompanying:

Q=set of states: qo (begin state), .qf (last state)

A=move likelihood network of n Xn probabilities of transitioning between any pair of n states (n=F+1)

-former likelihood (label arrangement)

O=succession of T perceptions (words) from a vocabulary V

–B=Succession of perception probabilities (likelihood of perception produced at state)

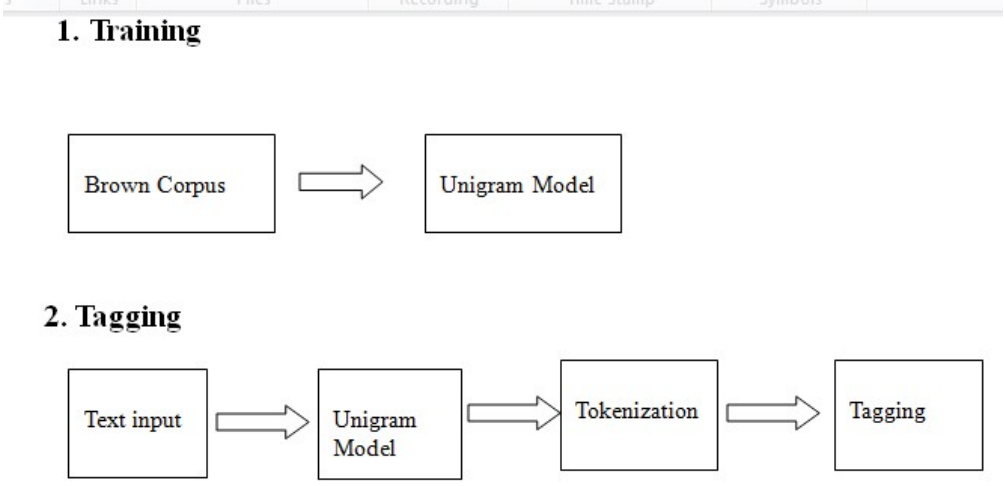
-Probability (word grouping)

Chapter 5

Analysis And Result

5.1 Part Of Speech Tagging using Unigram Model

Figure 5.1: Unigram

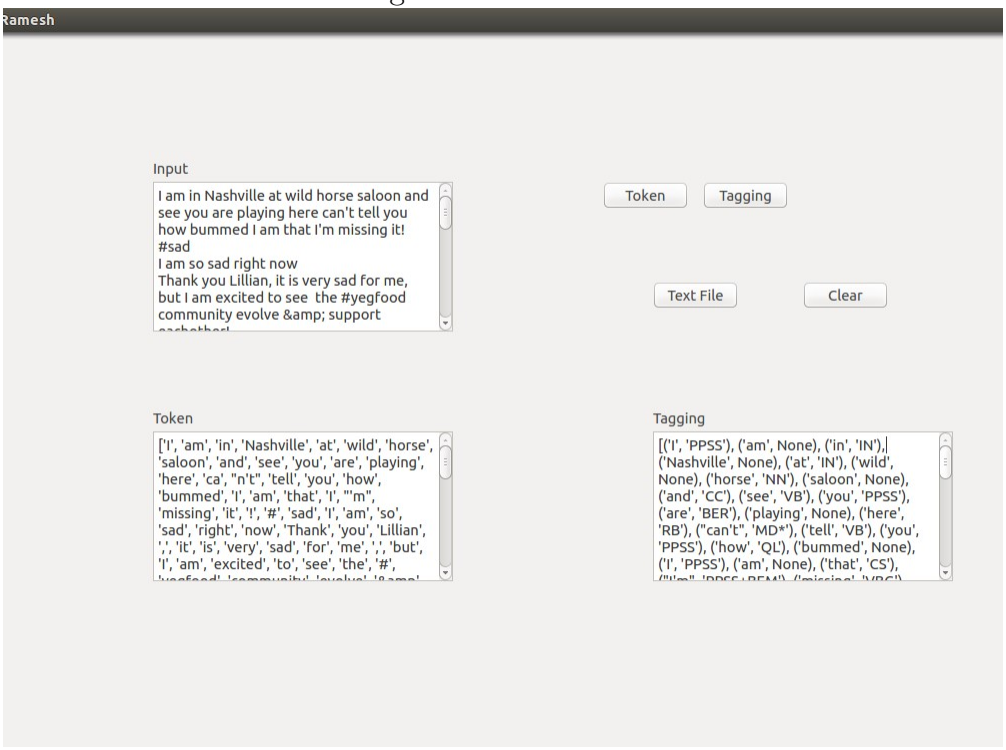


In POS tagging with Unigram model, we divided the whole process into two phases, namely Training and Tagging. Training: In training we are taking brown corpus as our training data. Brown corpus is the standard tagged data available. Next we need to train the unigram tagger with the brown corpus data. Unigram tagger is available in the NLTK (Natural Language Processing Tool Kit). We are using this unigram tagger and training it with brown corpus

data. Once the brown corpus data is converted into the required form which can be accepted by the NLTK unigram tagger, the tagger can be trained.

The performance of the tagger is more if we use more data for training, but it may effect the training time. Testing: Now to test the performance and accuracy of the trained tagger we need to test it with some data. Now take collections of sentences and input to the tagger. The tagger will tag the input and return the result in a list of tuples. Each tuple consists of the word and its tag. This input and output can be handled by the application shown in fig 5.1This application is made with Qt to work in Linux.

Figure 5.2: Tool Box



Above tool box shows text data is taken in input and this data is tokenized and tagged.

Tagging

[('I','ppss'),('am',none),('in','IN'),('Nashvile',none),('at','IN'),('wild',none),('horse','N

Now checking the performance of the tagging with the increase in size of the test data. The output simulation is shown in the figure below. From the

table 5.1. we can infer that with the increase in size of the input, the tagging time also increases linearly.

Table 5.1: TAGGING TABLE:

No	TAG	DESCRIPTION
1	PPSS	It indicates personal pronouns
2	NONE	Default tagger
3	IN	Preposition
4	NN	Noun
5	CC	Conjuction
6	VB	Verb
7	BER	Verb "to do", presentense, 2nd person singular or all person plural, negated
8	RB	Adverb Ex: only, often, also, there, etc..
9	QL	Qualifier Ex: well, less, very, most, so, etc

Figure 5.3: Simulation Diagram

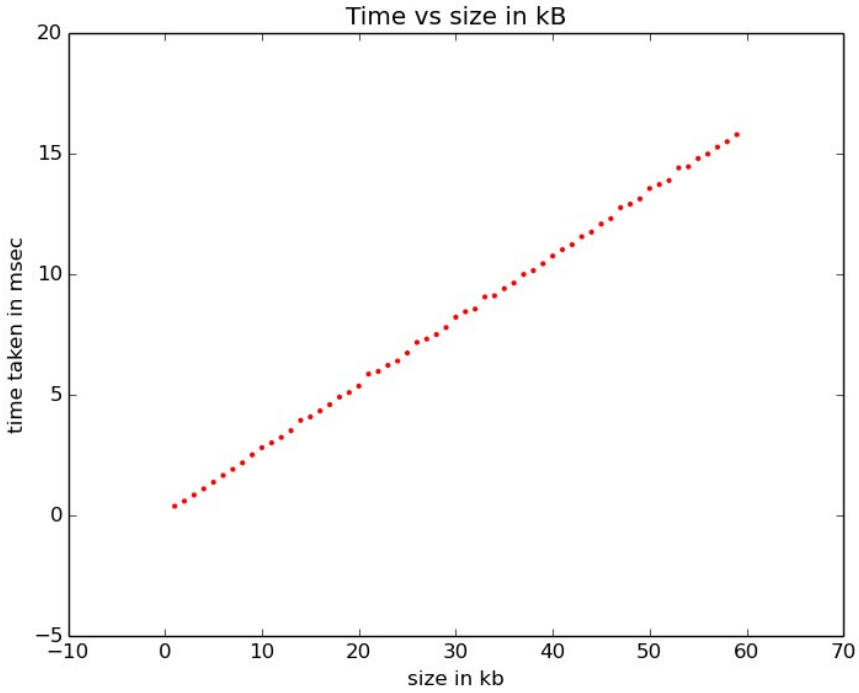
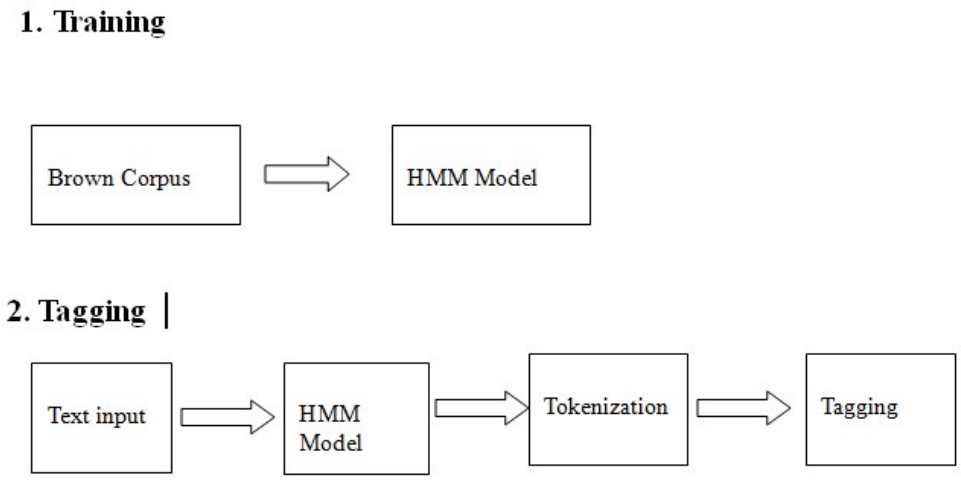


Figure 5.4: Part Of Speech Tagging using Hidden Markov Model



5.2 Part Of Speech Tagging using Hidden Markov Model

In POS tagging with Hidden Markov model, we divided the whole process into two phases, namely Training and Tagging. Training: In training we are taking brown corpus as our training data. Brown corpus is the standard tagged data available. Next we need to train the HMM tagger with the brown corpus data.HMM tagger is available in the NLTK (Natural Language Processing Tool Kit). We are using this HMM tagger and training it with brown corpus data. Once the brown corpus data is converted into the required form which can be accepted by the NLTK Hmm tagger, the tagger can be trained. The performance of the tagger is more if we use more data for training, but it may affect the training time. Testing: Now to test the performance and accuracy of the trained tagger we need to test it with some data. Now take collections of sentences and input to the tagger. The tagger will tag the input and return the result in a list of tuples. Each tuple consists of the word and its tag.

Now the accuracy of both HMM tagger and the Unigram tagger are calculated to check for the better performance. Performance can be measured with accuracy with the increase in number of tokens to tag. The result of

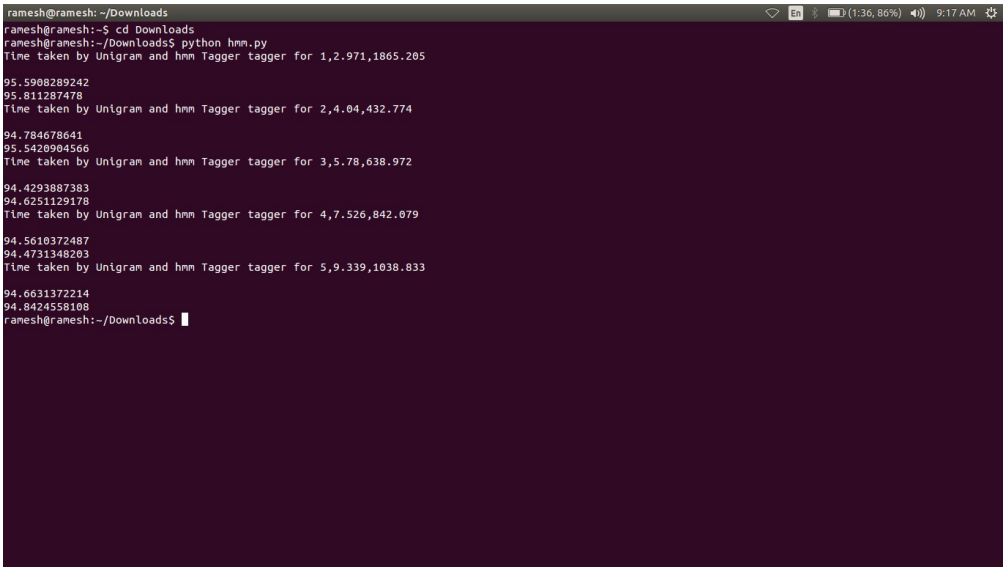


Figure 5.5: Performance of Unigram and HMM

simulation can be shown in the below graph. From the graph we can infer that the HMM tagger is more accurate than the Unigram tagger and also with the increase in the number of tokens the accuracy decreases and is almost constant after a certain number of the input tokens

5.3 Performance of Unigram and HMM

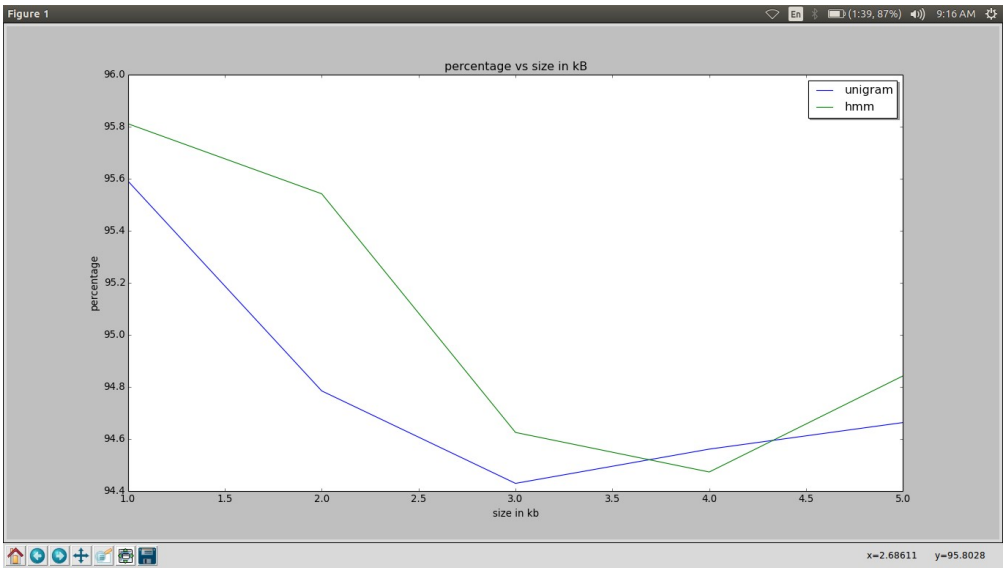


Figure 5.6: Graphical Comparison of Unigram and HMM

Chapter 6

Conclusions

This thesis is about the study of Parts of Speech tagging, particularly Unigram tagger and the Hidden Markov Model tagger. We presented the working principle of both the taggers and the by using the NLTK tool kit we have succeeded in developing a tool to tag the English sentence using both the taggers. The performance of each tagger with various constraints is also calculated and plotted. In our work it is found that both the taggers with the increase in size of the input, their performance varied linearly and with the increase in number of tokens, the accuracy of the HMM tagger is more when compared to the unigram tagger. Even though the tagging HMM tagger takes more time its accuracy is more than unigram tagger.

References

[1] G. K. Karthik ,K. Sudheer, and A. Pvs *Comparative Study of Various Machine Learning Methods for Telugu Part of Speech Tagging*,In Proceeding of the NLPAI Machine Learning Competition, 2006.

[2] L. V. Guilder *Automated Part of Speech Tagging: A Brief Overview*, Handout for LING361, Georgetown University, Fal, 1995.

[3] C. M. Kumar *Stochastic Models for POS Tagging*,IIT Bombay, 2005

[4] E. Brill *Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging*, Department of Computer Science Johns Hopkins University, 1996

[5] B. Pang and L. Lee *A sentimental education Sentiment analysis using subjectivity summarization based on minimum cuts*,Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL, 2004.

[6] J. Wiebe and E. Riloff *Creating subjective and objective sentence classifiers from unannotated texts*, in Computational Linguistics and Intelligent Text Processing. Springer, 2005, pp. 486–497