Analysis of Part of Speech Tagging

P. S. Patheja
HOD, M.Tech Dept.
BIST, Bhopal, India

Akhilesh A. Waoo
Assistant Professor,
M.Tech Dept
BIST, Bhopal, India

Richa Garg
Student, M.Tech.
BIST, Bhopal, India

ABSTRACT

In the area of text mining, Natural Language Processing is an emerging field. As text is an unstructured source of information, to make it a suitable input to an automatic method of information extraction it is usually transformed into a structured format. Part of Speech Tagging is one of the preprocessing steps which perform semantic analysis by assigning one of the parts of speech to the given word. In this paper we had discussed various models of supervised and unsupervised technique shown the comparison of various techniques based on accuracy, and experimentally compared the results obtained in models of Supervised Condition Random Field and Supervised Maximum Entropy model. We had deployed a model of part of speech tagger based on Hidden Markov Model approach and had compare the results with other models. Also we had discussed the problem occurring with supervised part of speech tagging.

General Terms

Supervised Technique, Unsupervised Technique, Part of Speech Tagging, Accuracy.

Keywords

NLP, CRF, MaxEnt, POS

1. INTRODUCTION

There is a wide range and focus areas in Human Language Technology (HLT). These include areas such as Natural Language Processing (NLP), Speech Recognition, Machine Translation, Text Generation and Text Mining. A natural language understanding system must have knowledge about what the words mean, how words combine to form sentences, how word meanings combine to from sentence meanings and so on.

Text documents are greatest source of information from which user extract information depending upon his interest [1] .So in order to extract meaning and relevant information in text document focus lies towards passage or sentence.

Retrieving relevant passage as compared to whole document helps in filtering out irrelevant information that improves accuracy [7].It runs into many stages, namely tokenization, lexical analysis, syntactic analysis, semantic analysis, pragmatic analysis and discourse analysis.

As text is an unstructured source of information, to make it a suitable input to an automatic method of information extraction it is usually transformed into a structured format. This preprocessing involves multiple steps namely sentence segmentation, tokenization, part of speech tagging, entity detection, relation detection [2].We in this paper are focusing on one of the preprocessing step i.e. part of speech tagging .

Parts of Speech Tagging is an approach to perform Semantic Analysis and include the process of assigning one of the parts of speech to the given word. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories.

Part of Speech Tagging has been broadly divided upon Supervised and Unsupervised Techniques having further classification of each type. In the remainder of this paper detailed classification of both Supervised and Unsupervised Techniques are described further stating the best techniques resulted based on accuracy achieved so far. We had then shown experimental results obtained for two best of art Part of Speech Tagging techniques based on their execution time. The results of a model based on Part of Speech tagger has been demonstrated which has been developed taking WordNet as a lexicon. Finally we had discussed the issues occurring in supervised system of tagging.

2. CLASSIFICATION OF PART OF SPEECH TAGGING

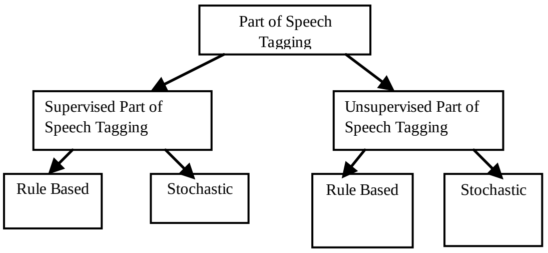
Part of Speech Tagging
Supervised Part of Speech Tagging
Unsupervised Part of Speech Tagging
Rule Based
Stochastic
Rule Based
Stochastic

Fig.1 Broad Classification of Part of Speech Tagging Techniques

Tagging in natural language processing (NLP) refers to any process that assigns certain labels to linguistic units. It denotes the assignment of part-of-speech tags to texts. A computer program for this purpose is called a tagger. Part of speech tagging includes the process of assigning one of the parts of speech to the given word. For example, the english word rust for instance is either a verb or a noun. Part of speech tagging can be categorized as follows:

2.1 Supervised and Unsupervised Taggings

Supervised Technique use a pre-tagged corpora (structured collection of text) which is used for training to learn information about the tagset, word-tag frequencies, rule sets etc. Supervised Classification mainly comprise of two phases i.e. training and prediction. During training different class labels are being generated by feature extractor, which convert each input value to a feature set or class label. So during training a pair of feature set and class label are fed into machine learning algorithm to generate model. During prediction phase, same feature extractor is used for generating predicted labels for unseen input or test set. Unsupervised Part of Speech (POS) tagging models do not require pre-tagged corpora. Operates by assuming as input POS lexicon, which consists of a list of possible POS tags for each word. [3].

2.2 Rule Based and Stochastic Techniques

Stochastic tagging is the phenomena, which incorporates frequency or probability, i.e. statistics.
Rule based techniques use contextual and morphological information to assign tags to unknown or ambiguous words. These rules are often known as context frame rules for example: If an ambiguous/unknown word X is preceded by a determiner and followed by a noun, tag it as an adjective.[8]
These rules can be either automatically induced by the tagger or encoded by the designer. Eric Brill designed the best-known rule-base part of speech tagger, which was the first one to be able to achieve an accuracy level comparable to that of stochastic taggers i.e 95%-97% [4].

3. CLASSIFICATION OF SUPERVISED AND UNSUPERVISED TAGGING TECHNIQUES

Supervised and Unsupervised tagging techniques can be classified into following categories.

3.1 Decision Tree Model

A decision tree is a predictive model with a tree structure that recursively partitions the training data set. Each internal node of a decision tree represents a test on a feature value, and each branch represents an outcome of the test. A prediction is made when a terminal node (i.e., a leaf) is reached. Tree tagger is able to achieve the accuracy of 96.36% on Penn Treebank better than of trigram tagger (96.06%). [12]

3.2 Condition Random Field Model

J.Lafferty explores the use of Condition Random Field (CRF) model for building probabilistic models and labeling sequence data. They are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. Conditional random fields (CRFs) for sequence labeling offer advantages over both generative models like Hidden Markov Model (HMM) and classifiers applied at each sequence position. CRFs don't force to adhere to the independence assumption and thus can depend on arbitrary, non-independent features, without accounting for the distribution of those dependencies [17].

CRF is defined as follows. Let **y** be a set of output variables that we wish to predict, and **x** be a set of input variables that are observed. For example, in natural language processing, **x** may be a sequence of words **x** = $\{x_t\}$ for t = 1……..T and **y** = $\{y_t\}$ a sequence of labels. Let G be a factor graph over **y** and **x** with factors $C = \{\varphi_c(y_c, x_c)\}$; where $\mathbf{x}_c$ is the set of input variables that are arguments to the local function φ_c, and similarly for $\mathbf{y}_c$. A conditional random field is a conditional distribution p that factorizes as follows $p(y|x) = \frac{1}{z(x)} \prod_{c\in C}\{\varphi_c(y_c, x_c)\}$; [14] Where z(x) is a Normalization factor over all state sequences for the sequence x and $z(x) = \sum_y \prod_{c\in C} \varphi_c(y_c, x_c)$; CRF achieves accuracy of 98.05% in close test and 95.79% in open test [13].

3.3 Hidden Markov Model

In Hidden Markov Models (HMM) state transitions are not observable. HMM taggers require only a lexicon and untagged text for training a tagger. Hidden Markov Models aim to make a language model automatically with little effort. Disambiguation is done by assigning more probable tag.
A Hidden Markov Model (HMM) consists of the states which correspond to the tags, it has an alphabet which consists of the set of words, the transition probabilities P ($\text{Tag}_i|\text{Tag}_i$-1) and the emission probabilities P($\text{Word}_i|\text{Tag}_i$).In HMM, for a given (word, tag) pair we have the probability:

$$P(w, t) = \Pi\, P(\text{Tag}_i|\text{Tag}_i\text{-}1) * P(\text{Word}_i|\text{Tag}_i). \qquad (1)$$

Different work carried out under HMM are that of (Merialdo, 1994), (Elworthy, 1994), (Banko and Moore 2004), (Wang and Schuurmans 2005) [2].
Maximum accuracy obtained is 95%-97% [5].

3.4 Maximum Entropy Model

Maximum Entropy Tagging thrives to find a model with maximum entropy. Maximum entropy is the maximum randomness. The outputs of the maximum entropy tagging are tags and their probabilities. Maximum entropy model specifies a set of features from the environment for tag prediction. In contrast to HMMs, in which the current observation only depends on the current state, the current observation in an MEM may also depend on the previous state. The term, maximum entropy here means maximum randomness or minimum additional structure. Best accuracy reported in maximum entropy model is by Stanford tagger of 96.9%. [17]

3.5 Clustering Model

This model focuses distributional properties and co-occurrence patterns of text (similar words occur in similar contexts) by computing context vectors for each word to cluster words together in groups, groups which can then be assigned Part of Speech tags or word classes as groups.
The key characteristics are how the context vectors are defined, size of the context vectors (number of dimensions),metric used to compute vector similarity (i.e. make clusters),and how the tags or word classes are induced on the clusters. (Schutze, 1995) and (Clark, 2000) had shown results in this Category of clustering model. Best accuracy is reported as 59.1%. [6]

3.6 Prototyping Model

Prototypes posses' better evaluation (since small size) and more meaning than clusters. In this model a few examples or prototypes are collected (one for each target tag) and then propagated across the corpus of unlabeled data. No lexicon is required in this model. A gradient based search with the forward-backward is used to maximize the log linear model parameters. Accuracy achieved in this model is 80.5%. [15]

3.7 Bayesian Model

Bayesian learning models for Part of Speech tagging integrates over all possible parameter values as compared to finding a parameter set which maximizes the probability of tag sequences given unlabeled observed data. Work done in Bayesian Model is shown by (Toutanova and Johnson, 2007), (Goldwater and Griffiths, 2007), (Johnson, 2007). [2] Accuracy achieved is 93.4%

3.8 Neural Networks

A neural network (NN) is an interconnected group of natural or artificial neurons that uses a computational model for processing data pairs of input feature and desired response where data pairs are input to the learning program.
Input features partition the training contexts into a number of overlapping sets corresponding to the desired responses.
Best accuracy achieved in neural network is 96.9% [17].

4. DATA SOURCES REFFERED DURING PART OF SPEECH TAGGING

Knowledge is a fundamental component of part of speech tagging. Knowledge sources provide data which are essential to associate senses with words. Knowledge sources can be divided into following types.

4.1 Machine Readable Dictionaries

MRD are dictionaries in electronic format which are most utilized resource for word sense disambiguation in English.WordNet encodes a rich semantic network of concepts and defined as a computational lexicon.

4.2 Corpora

A corpus (plural *corpora*) or text corpus is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. Examples include BNC, SemCor e.t.c.

5. COMPARISON OF DIFFERENT MODELS OF SUPERVISED AND UNSUPERVISED TECHNIQUE

Comparative results have been shown in table 1 for different models of Part of Speech Tagging Technique based on data obtained from different reference papers and sources and correspondingly best accuracy results had been demonstrated by two supervised tagging technique i.e CRF and Maximum Entropy model.

TABLE 1. Comparative Performance Results

Method	**Accuracy in %**
Decision Tree	93.36
Max Entropy	96.97
HMM	95-97%
CRF	95.79-98.05%
Clustering	59.1%
Prototyping	80.5%
Bayesian	93.4%
Neural Network	93.4%
Rule Based	95-97%

6. EXPERIMENTAL RESULTS

6.1 Condition Random Field Model and Maximum Entropy Model

A maximum entropy based tagger has been proposed in (Ratnaparkhi, 1996).The tagger learns a log linear conditional Probability model from tagged text, using maximum entropy method.We had used Stanford part of speech tagger, which is an extension of the paper Ratnaparkhi [10] further incorporating log linear concept in maximum entropy model. This tagger uses Penn Treebank comprising a set of around 48 tags for tagging. The tagger has three modes tagging, training, and testing. Tagging allow us using a pretrained model to assign part of speech tags to unlabeled text. Training allows us saving a new model based on a set of tagged data. Testing allows us to see how well a tagger performs by tagging labeled data and evaluating the results against the correct tags. We had experimented the results with varying number of tokens and the correspondingly execution rate achieved. As stated in [17] the best accuracy reported in Maximum Entropy model ranges from 96.97%-97%.Correspondingly the performance of tagger in terms of efficiency is demonstrated in Table 2.
CRF tagger had been helping in dealing with the label bias problem present in Maximum Entropy Markov Model [16].CRF model address the problem by using a single exponential model for entire label sequence given a observation sequence. Table 3 shows the performance of CRF model based tagger in terms of the efficiency achieved for same dataset as used for Maximum Entropy model. When demonstrating tagging results with Condition Random field model we had used Penn Treebank tagset.CRF tagger is unable to demonstrate accurate results for small number of tokens.

Table 2 .Results obtained for Stanford Tagger stating the results obtained for Maximum Entropy model

S.No	**Number of Tokens**	**Execution Rate per sec**	**Execution Time (in sec)**
1	04	85.11	.0469
2	08	170.21	.0470
3	32	680.85	.0470
4	0513	2052.00	0.25
5	0989	2108.74	0.469
6	3896	2770.98	1.406
7	15584	5167.11	3.015
8	31168	5334.25	5.842

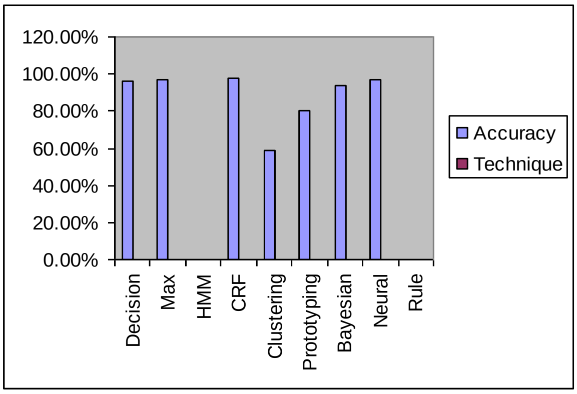

Fig 2. Graphical representation of different tagging Techniques based on accuracy obtained from various sources

6.2. Proposed Work: New Tagger Developed Deploying WordNet as Lexicon

We had developed a model of part of speech tagger using WordNet as a computational lexicon the tagger derives a probability formulae where w = Word, t = Tag

$$P (w \mid t) = \sum P (wi \mid ti) + \sum (ti \mid ti\text{-}1) \qquad (2)$$

The results of proposed tagger based accuracy is represented in Table 4. and correspondingly its graphical representation is demonstrated in Figure 3 .

Table 3. Results obtained for Condition Random Field Model

S. No	Number of Tokens	Execution Time (in sec)
1	04	0.0
2	08	0.0
3	32	0.0
4	0513	0.031
5	0989	0.063
6	3896	0.203
7	7792	0.375
8	15584	0.781
6	31168	1.531

Table 4. Results obtained for new developed Part of Speech Tagger

S. No	No. of Token	Execution time (sec)	Accuracy (%)
1	4	0.022	88.46
2	8	0.031	92.10
3	12	0.040	86.46
4	16	0.046	83.78
5	32	0.399	82.48

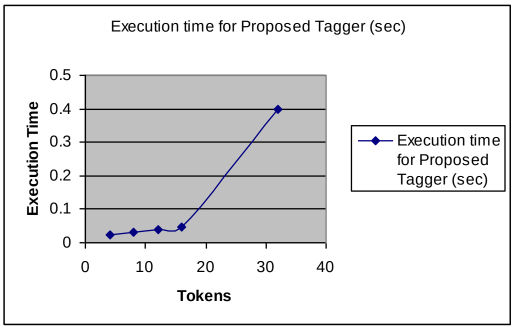

Fig. 3 Graph shows efficiency results for proposed tagge

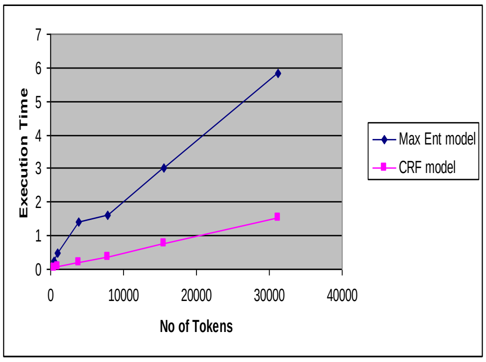

Fig. 4. Graph shows CRF model had shown results with improved performance in terms of execution time as compared to Max Entropy (Max Ent) model.

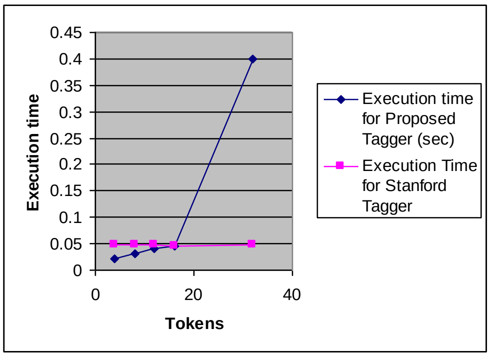

Fig. 5. Graph shows proposed tagger achieve results With improved performance in terms of execution time as Compared to Max Entropy (Max Ent) model.

7. DISCUSSION

Condition Random Field (CRF) based model attains good performance results as compared to Maximum Entropy Model As shown experimentally in Fig.4. Also the model of tagger using WordNet as lexicon has demonstrated sufficiently good efficiency when compared with Maximum Entropy model as shown in Fig. 5.

Though supervised technique had shown good performance results in terms of accuracy, yet it suffers from the problem of data sparcity.Data Sparcity is the issue in which words appearing in test set are unavailable in test set due to large size of dictionaries. Research is going on to solve this issue of data sparcity with the help of CRF model.

8. LIMITATION

The proposed tagger have demonstrated good results for very small number of tokens but due to some memory issues during coding fail to demonstrate better results for large number of tokens.

Also a limitation lies with the CRF tagger i.e. this tagger fails to give results for very small number of tokens due to which we have shown comparative results with Maximum Entropy based tagger only.

9. CONCLUSION

After comparing the experimental results of both model it is found that for a dataset of around 31k tokens the average execution time obtained for Maximum Entropy model is 2.0675 sec and for Condition Random Field model is 0.49733 sec respectively. Thus it is proved that CRF model achieve better performance results both in terms of accuracy and execution time(as shown in Fig 4.) than Maximum Entropy model. Also the implemented tagger using WordNet as a lexicon has demonstrated better performance results for small number of tokens than Maximum Entropy based tagger i.e. for a dataset of around 32 tokens the average execution time obtained for Maximum Entropy model is .0496 sec and and for new tagger is 0.0357 sec respectively (as shown in Figure 5) Further a scope of handling the data sparcity issue provide a vast area of research in the field of tagging.

10. REFERENCES

[1] Zhongqiang Huang, Vladimir Eidelman, Mary Harper, June 2009. Association for Computational Linguistics Proceedings of NAACL HLT.

[2] Roberto Navigli, University `a di Roma La Sapienza, Publication date: February 2009,ACM Computing Surveys, Vol. 41, No. 2, Article 10.

[3] William P. Headden III, David McClosky, Eugene Charniak, August 2008, Proceedings of the 22nd International Conference on Computational Linguistics.

[4] Garnett Wilson and Malcolm Heywood, June 25–29, 2005, Washington, DC, USA.Copyright 2005 ACM 1-59593-010-8/05/0006.

[5] Sujith Ravi and Kevin Knight, 2009 ACL and AFNLP.

[6] Valentin I. Spitkovsky, Hiyan Alshawi, Angel X. Chang, Daniel Jurafsky, Unsupervised Dependency Parsing without Gold Part-of-Speech Tags.

[7]Genoveva Galarza Heredero, Subhadip Bandyopadhyay, Arijit Laha, Copyright 2011, ACM 978-1-4503-0750-5/11/03.

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

[9][14]Eric Brill, A Simple Rule Based Part of Speech Tagger ACLTrento Italy, 1992,Proceedings of the Third Conference on Applied Computational Linguistics

[10]Kristina Toutanova, Christopher D. Manning, Stanford, CA 94305–9040, USA.

[11] Sharon Goldwater and Mark Johnson,2005 ACL Representation Bias in Unsupervised learning of Syllable Structure.

[12] Helmut Schmid, 1994 Probabilistic Part of speech Tagging Using Decision Trees, Manchester, UK.

[13] Xiaofei Zhang,Heyan Huang,Zhang Liang , IEEE-Nov 2009,The application of CRF in Part of Speech Tagging , 9778-0-7695-3752-8.

[14]Charles Sutton, Andrew McCallum, Khashayar Rohanimanesh ,Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data, Journal of Machine Learning Research Published 3/07, 8 (2007) 693-723.

[15]Aria Haghighi, Dan Klein, Prototype-driven learning for sequence models,06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics .

[16] Ryan T. K. Lin, Hong-Jie Dai, Yue-Yang Bow,Justing Lian-Te Chiu ,Richardg Tzon-Han Tsa , December 2009,Using conditional random fields for result identification in biomedical abstracts,Published in: Journal Integrated Computer-Aided Engineering.

[17] http://nlp.stanford.edu.