Research on Machine Learning Techniques for POS Tagging in NLP

Aparna Bulusu, Sucharita V

ABSTRACT---Natural Language Processing is an emerging area with applications like speech recognition, sentiment analysis, question answering systems, chat bots and the like. Current research is making heavy use of machine learning techniques for NLP Tasks Machine Learning is a subset of Artificial Intelligence with techniques based heavily on statistical and mathematical concepts. This paper attempts to list out the major categories of tasks under Natural Language Processing, and understand the commonly used machine learning techniques for the said tasks. Empirical tests have been performed to validate the findings of the literature survey and results are discussed.

Keywords—Natural Language Processing, Machine Learning, Brown Corpus, Classifiers, NLTK

I. INTRODUCTION

Natural Language Processing can be defined as the automatic processing of natural language either in speech or textual form, through software. Once considered a part of computational Linguistics, NLP research is now an independent and heavily researched area due to the improvements in computing power and availability of large annotated databases. Natural Language Processing consists of many challenging tasks like POS Tagging, Chunking, Semantic Role Labeling, Tokenization, Stemming, Word Sense Disambiguation, Named Entity Recognition etc. [1]

A wide variety of applications are being driven by NLP. Some common applications include information retrieval tasks, automatic summarization of text, sentiment analysis, Text Classification, Question Answering Systems, chat bots for customer service, Automated voice controlled services like Alexa and Siri, gathering business intelligence, automated machine translation, text-to-speech and vice versa, natural language understanding and language generation , text analytics etc. Due to the immense potential of this field, a lot of current research is focused on finding appropriate solutions for the various subtasks within NLP

II. AN OVERVIEW OF MACHINE LEARNING

Machine Learning can be thought of as the ability of machines / computers to learn a specific task without explicit programming. Machine learning Techniques have been broadly classified as supervised learning methods, unsupervised learning methods, reinforcement based methods and currently a new development in the field is deep learning.

Supervised Learning methods are based on the concept of training a model using existing data sets that are classified with class labels and then implementing this model on unseen data to predict results. Most Classification tasks are solved through supervised techniques. Most common supervised machine learning techniques include support vector machines, rule based classifiers, Bayesian models, hidden markov models, models based on maximum entropy, decision trees, perceptron's and artificial neural networks.[2]

Unsupervised techniques are more suitable for tasks that involve clustering or grouping related data. Many clustering approaches are used like K Means, K Nearest Neighbours, Single and Complete link Clustering, Agglomerative Hierarchical Clustering, latent semantic analysis, Apriori algorithm for association rule learning etc.Semi supervised learning combines both supervised and unsupervised techniques to generate results. Reinforcement Learning is based on the concept of an agent performing actions within an environment to move from one state to another and trying to maximize the reward for its actions. This is an iterative process where the model is continually learning from its environment.

III. LITERATURE REVIEW – COMMON NLP TASKS

POS Tagging

Part Of Speech Tagging: This is considered as one of the initial pre-processing stages in NLP. Also multiple models have been developed for POS Tagging with close to 97% tagging accuracy. The POS Tagging task involves identifying the right part of speech for each individual token and tagging it accordingly. Tagging on English texts has been carried out extensively using many machine learning tehniques. A Maximum Entropy model that uses probability methods has been shown to reach over 96% tagging accuracy. [3]

The Trigrams N Tags Tagger makes uses of statistical methods and its implementation through HMM's has been proven to be an efficient technique for tagging.[4]. Support Vector Models are also commonly used for POS Tasks. [5] reports that support vector machines are better than HMM's in certain cases to perform POS Tagging. They make use of contextual and sub string information of preceding and succeeding terms which can be effectively combined in SVM's and guesses POS tags of unknown words.

Revised Manuscript Received on June 10, 2019.
AparnaBulusu,Research Scholar, Dept of CSE K L UniversityVaddeswaram, Vijaywada, A.P, India.(aparnabulusu79@gmail.com)
Dr Sucharita V Professor, Dept of CSE NarayanaEngg College Gudur, Nellore,A.P, India. (jesuchi78@yahoo.com)

Hidden Markov Models are also suitable for carrying out POS Tagging. [6] reports that morphologically rich languages with a high number of POS tags can't be handled well with regular methods like SVM's and TnT. Ensemble learning is a technique in which multiple machine learning techniques are stacked together. [7] reports that an ensemble tagger based on multiple decision trees provides good tagging accuracy.

Chunking

This is a NLP subtask that usually follows the POS Tagging stage. It involves combining multiple individual tokens into related chunks like noun phrases through the use of regular expressions. Chunking is especially useful to extract information regarding names of entities like people, corporations, locations etc. [8] have found that support vector machines are extremely suitable for the chunking process as they achieve good generalization performance. Datasets used in NLP are typically known for having a high number of features and SVM's are very suitable for feature reduction. [9] treat chunking as a sequential prediction problem and make use of a generalized form of the winnow algorithm to achieve text chunking with great accuracy. Literature also suggests that use of semi supervised learning methods is effective in text chunking.[10] make use of a unified deep learning architecture to effectively handle most of the NLP tasks mentioned above. They make use of the concept of multi task learning wherein the deep neural net model is trained to handle multiple tasks jointly. This approach has resulted in most of the tasks being accomplished with good accuracy.

Word Sense Disambiguation

Natural language is considered to be highly ambiguous i.e. A term or word can take on different meanings within different contexts and correctly identifying its semantic and grammatical meaning is a difficult task. WSD requires having extensive lexical and syntactic knowledge about a language and use of preexisting resources like Thesauri, Machine readable dictionaries and ontologies, annotated corpora and collocation resources[11]. Decision Lists and Trees, Neural Networks, SVM's , Instance based learning and ensemble methods are the supervised techniques mostly employed for performing Word Sense Disambiguation. Unsupervised methods include context based clustering, word based clustering, using co-occurrence graphs. Graph based algorithms have also been successfully applied for corpus based disambiguation of word senses. [12]

Sematic Role Labeling

This is an advanced NLP task that involves trying to figure out responses to questions like 'who' , 'what', 'where', 'when' etc. Semantic Role Labeling has applications in information extraction, summarization, translation etc.[13]

[14] states that most common techniques used for semantic role labeling include support vector machines with kernels, Maximum entropy methods, Decision Trees, Conditional Random Fields and Memory Based Learning Techniques

Named Entity Recognition

Named entities are defined as phrases that contain the names of people, Locations, Organizations etc. The most commonly applied technique for Named Entity recognitionwas Maximum Entropy Model. [15] Other methods that give good results include use of Hidden Markov Models , support vector machines, conditional random fields and algorithms like Adaboost.

Other tasks within NLP Domain are tokenization, Parsing, syntactical analysis, stemming etc. Machine learning techniques are available for nearly all these tasks especially for languages like English that have a large collection of annotated corpora available freely online.

IV. METHODOLOGY & RESULTS

The current work is focused on applying the ML methods to one specific NLP task - POS Tagging. The following methodology was used

Brown Corpus is a freely available tagged corpus available online for the purpose of training and testing our classifiers. It is considered as one of the most important tagsets on which subsequent corpus like Penn TreeBank were constructed[16]

NLTK - the Natural Language Toolkit, is a suite of open source program modules, tutorials and problem sets, that are used for natural language processing. NLTK provides interfaces to annotated corpora and also has multiple implementations of various machine learning algorithms for achieving many NLP tasks[17]

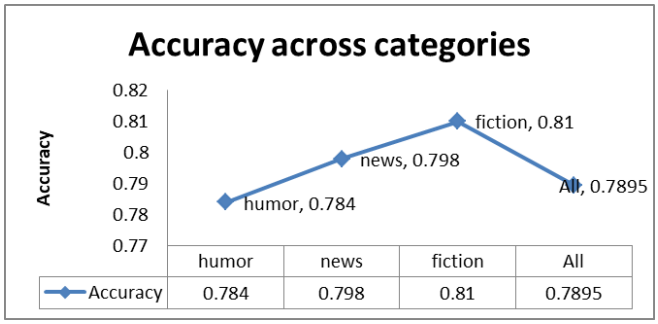
The current work was carried out on a desktop computer with intel i7 Core Processor with 8 GB RAM running Python Version 3.7 . Various categories of text from Brown Corpus were tagged with python programs using NLTK framework. Various parameters were varied like Size of Training Vs Test Sets, Categories of Text and Classifier algorithms and the corresponding POS Tagging accuracy levels were recorded.

1. Results & Discussions

The Brown Corpus consists in total of around 1 million words across multiple categories [18] A base tagger was created that used the information of suffixes of words upto length 2 in oreder to determine a word's tag. Each category of text was divided into a training set with 90% of data and remaining was used as attest set to examine tagging accuracy. The navebayes classifier from NLTK3 was used to train the tagger. Results of varying the category are as follows:

Category	Length of corpus (Sentences)	Training set size	Test set Size	POS tagging accuracy (Using Naive Bayes Classifier
humor	1053	948	105	0.784
news	4623	4161	462	0.798

fiction	4249	3825	424	0.81
All (Complete Corpus)	57340	51606	5734	0.7895



It is observed that tagging accuracy is slightly better for the 'fiction' category.

II. Effect of Varying Test and Training Sizes

The news category from Brown corpus was chosen to understand the impact of varying the training and test size. 90% of the corpus (under news category) was assigned as the training set. A tagger with trained to extract word features based on suffix and prefix information, and was used to tag the training set using a naïve bayes classifier. This tagger was then used to evaluate the accuracy against the test set. The same tagger was implemented multiple times by varying the sizes of training and test sets as (80,20), (70,30), (60, 40) and finally 50% of corpus as training data and 50% of data as test data. The following levels of accuracy were noted.

Training Set Size	Test Set Size	Tagging accuracy
4161	462	0.798
3699	924	0.7937
3237	1386	0.7811
2774	1849	0.7769
2312	2311	0.7742

III. Effect of Tagsets used in training

The default tagset that is used by NLTK for tagging purposes is the 'Universal tag set' . However multiple tagsets with varying number of tags are available like Penn Treebank set, Claws5 and Brown tagset. The same corpus (News category from Brown Corpus) was tagged according to universal and brown tagsets and top 10 frequent kinds of tags with their frequency were reported. The results clearly demonstrate that accuracy is dependent on our choice of tag set

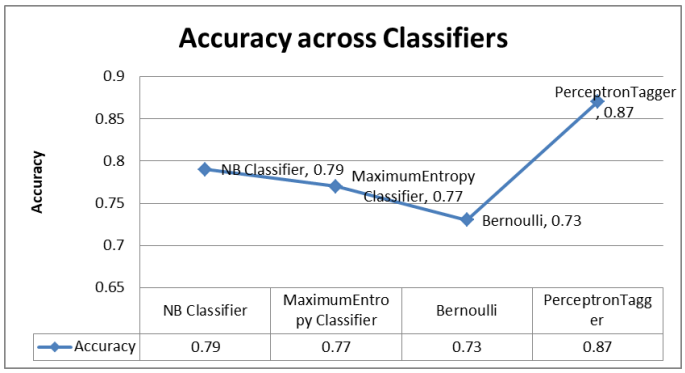
Using the default 'Universal' Tag set	**Using the 'Brown' Tag set**
[('NOUN', 30654), ('VERB', 14399), ('ADP', 12355), ('.', 11928), ('DET', 11389), ('ADJ', 6706), ('ADV', 3349), ('CONJ', 2717), ('PRON', 2535), ('PRT', 2264), ('NUM', 2166), ('X', 92)]	('NN', 13162), ('IN', 10616), ('AT', 8893), ('NP', 6866), (',', 5133), ('NNS', 5066), ('.', 4452), ('JJ', 4392), ('CC', 2664), ('VBD', 2524), ('NN-TL', 2486), ('VB', 2440)

IV. Effect of varying underlying algorithms

The tagger program was altered to use algorithms from scikit learn package – BernoulliNB classifier, Maximum Entropy Classifier and finally the same corpus was tagged by the built in PerceptronTagger available in NLTK. The following results were reported:

NaiveBayesClassifier	0.79
MaximumEntropy Classifier (with 100 iterations)	0.77
BernoulliNB Classifier	0.73
Using a PerceptronTagger	0.87

Perceptron based Tagger achieved the highest level of accuracy among all variants.



V. CONCLUSION

A review of existing literature suggests that a large number of machine learning techniques are being utilized for solving NLP Tasks. It has been observed that minor differences like content category, choice of tag set, size of training sets can have an impact on accuracy of the POS taggers. Choice of ML algorithms for performing the actual tagging is very important and has a significant impact on accuracy levels. Further work will focus on creating a gold standard tagger by using ensemble methods that yields comparable accuracy levels for all kinds of text.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12, no. Aug (2011): 2493-2537.
2. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.
3. Ratnaparkhi, Adwait. "A maximum entropy model for part-of-speech tagging." In Conference on Empirical Methods in Natural Language Processing. 1996
4. Brants, Thorsten. "TnT: a statistical part-of-speech tagger." In Proceedings of the sixth conference on Applied natural language processing, pp. 224-231. Association for Computational Linguistics, 2000.
5. Nakagawa, Tetsuji, Taku Kudo, and Yuji Matsumoto. "Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines." In NLPRS, pp. 325-331. 2001.
6. Schmid, Helmut, and Florian Laws. "Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging." In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pp. 777-784. Association for Computational Linguistics, 2008.
7. Marquez, Lluis, Horacio Rodriguez, Josep Carmona, and Josep Montolio. "Improving POS tagging using machine-learning techniques." In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999.
8. Kudo, Taku, and Yuji Matsumoto. "Chunking with support vector machines." In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, pp. 1-8. Association for Computational Linguistics, 2001.
9. Zhang, Tong, Fred Damerau, and David Johnson. "Text chunking based on a generalization of winnow." Journal of Machine Learning Research 2, no. Mar (2002): 615-637.
10. Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." In Proceedings of the 25th international conference on Machine learning, pp. 160-167. ACM, 2008.
11. Navigli, Roberto. "Word sense disambiguation: A survey." ACM computing surveys (CSUR) 41, no. 2 (2009): 10.
12. Bender, Oliver Agirre, Eneko, and Philip Edmonds, eds. Word sense disambiguation: Algorithms and applications. Vol. 33. Springer Science & Business Media, 2007
13. Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. "Semantic role labeling: an introduction to the special issue." (2008): 145-159.
14. Carreras, Xavier, and Lluís Màrquez. "Introduction to the CoNLL-2005 shared task: Semantic role labeling." In Proceedings of the ninth conference on computational natural language learning, pp. 152-164. Association for Computational Linguistics, 2005.
15. Franz Josef Och, and Hermann Ney. "Maximum entropy models for named entity recognition." In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 148-151. Association for Computational Linguistics, 2003.
16. Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993).
17. Loper, Edward, and Steven Bird. "NLTK: the natural language toolkit." arXiv preprint cs/0205028 (2002).
18. Kucera, Francis. "WN (1967). Computational analysis of present-day American English." Providence Brown UP.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication