A Survey on Techniques in NLP

Nihar Ranjan
Assistant Professor
Department of Computer Engineering,
Sinhgad Institute of Technology and Science,
Savitribai Phule Pune University

Kaushal Mundada
UG Student
Department of Computer Engineering,
Sinhgad Institute of Technology and Science,
Savitribai Phule
Pune University

Kunal Phaltane
UG Student
Department of Computer Engineering,
Sinhgad Institute of Technology and Science,
Savitribai Phule
Pune University

Saim Ahmad
UG Student
Department of Computer Engineering,
Sinhgad Institute of Technology and Science,
Savitribai Phule
Pune University

ABSTRACT
The field of natural language processing (aka NLP) is an intersection of the study of linguistics, computation and statistics. The primary goal of NLP is automated understanding of the semi-structured language that humans use. This study stems application in diverse fields like semantic analysis, summarization, text classification and the like. The domain natural language processing is a fledgling domain with no concrete indication of when it will mature. Compared to well established domains like "Study of Algorithms", NLP is yet in its emerging period and hence there's dearth of a concise piece of literature that elaborates on the phases of NLP and lists different techniques that can be adapted. NLP borrows heavily from foundational subjects of study like statistics, probability theory and theory of computation. In this paper, we describe three phases of natural language processing namely, language modelling, parts-of-speech tagging and parsing, outlining the approaches used that can be used.

Keywords
NLP, Language Modelling, Parsing, POS tagging, HMM

1. INTRODUCTION
The study of Language, ability to speak & write and communicate is one of the most fundamental aspects of human behaviour. As the study of human-languages developed the concept of communicating with non-human devices was investigated. This is the origin of natural language processing (NLP). The idea of natural language processing is to design and develop a computer systems that can analyse, understand and synthesis natural human-languages. Natural language falls within the domain of artificial intelligence with the goal of understanding and creating meaningful expressions in human-language. There are many applications of natural language processing developed over the years such as speech recognition, language translation, information retrieval, text summarization and the like. Before we dive into details, we must first consider the phases of NLP, or the pipeline through which a sentence goes before a parse tree of that sentence is built. NLP has several phases depending on the application but here, we will limit our discussion to the three phases namely, language modelling, parts-of-speech tagging and parsing.

2. OVERVIEW OF PHASES
The preliminary goal of any NLP application is to generate a parse tree for a sentence belonging to the set of that language. For creating a parse tree however, one needs to know the class to which all the words in the sentence belong that is whether a word is an adjective or a verb or something else. To correctly identify the class to which the particular word belongs to, we rely on the language model. Hence, the above stated actions are in reverse chronology and depend on each other as illustrated in the diagram. Note that the diagram is specific to the approach specified in this survey that is statistic language modelling, POS tagging and parsing. Certain approaches like neural networks, may not confirm to this chronological sequence.

One of the inherent problems that raises its head several times in NLP is the problem of ambiguity. Researchers have to deal with ambiguity in almost every phase of processing. For example, in POS tagging, consider the word "can": it could be classified as a modal verb because it is an ability to do something and it can also be classified as a noun because it can be a container that holds something. There are similar problems in the other phases.

Historically, the language processing applications worked by creating a rule based software that examined the structure of sentence to see if it fits the structure specified. Rule based approaches soon become unmanageable for large rules. With just over a hundred rules, the interaction between these rules becomes overly complex. These approaches were soon rendered useless by the sheer amount of data and rules that are applicable. Recent methods however, employ methods that take advantages of the deluge of data available for the purpose of training the language models. In other words, recent approaches to language processing make use of the data driven approaches to attain the goals of understanding language. These data driven strategies make up the statistical revolution of NLP. We will take a look at some of the statistical methods and discuss the results of alternative methods.

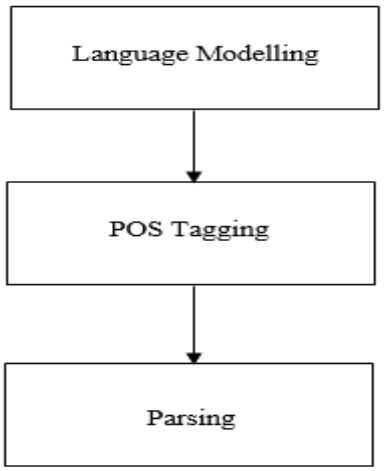

Figure 1

For the discussion, we are considering the following phases.

2.1 Language modelling

Language modelling is the art of making a probabilistic model of the language that is used by later stages o the language. This model is strictly statistical in the sense that it ignores the underlying meaning of the sentences and focuses on developing a probability distribution of the specified language.

2.2 POS tagging

Parts of speech tagging is the process of classifying the word in its context. It uses the probability model constructed in the previous section along with additional parameters to classify a word into its class.

2.3 Parsing

Parsing involves the construction of parse tree to understand the relation of different components of the sentence. This is especially important in resolving ambiguity. The parsing model uses a context free grammar along with probabilities associated with each rule to derive the parse tree of the sentence.

3. METHODOLOGIES' ANALYSIS

In this section, we will look at the available methods to perform the phases mentioned in the previous section. One of the methods is detailed in the process and results from different processes are stated.

3.1 Language Modelling

A statistical language model is simply a probability distribution over all possible sentences S in a language.[5] In other words, statistical language modelling only calculates the probability distribution of sentences without taking into account the semantics of the sentence. There are a number of ways to model a language such as n-gram models, decision tree model, linguistically motivated models, exponential models, adaptive models.

Here we will touch upon how n-gram models in language modelling work. N-gram models are the staple of language modelling process and the most widely used in speech recognition systems. N-gram models are based on the hidden Markov chain of order. A Markov chain is similar to conditional probability but with the assumption value changing according to the order of the Markov chain.

Consider random variables X_1, X_2, X_3 that can take on values x_1, x_2 and x_3. The probability of X_1, X_2, X_3 taking on values of x_1, x_2 and x_3 is given by

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

$$P(X_1 = x_1 | X_2 = x_2, X_3 = x_3).$$

$$P(X_2 = x_2 | X_3 = x_3).$$

$$P(X_3 = x_3)$$

But with the Markov assumption of degree one, it becomes the unigram model, with the Markov assumption of degree two, it becomes the bigram model. An n-gram model is conditioned on the previous terms.

To put it more succinctly in a mathematical formula, a bigram model will look like this

$$P(X_i = x_i)$$

$$= \sum_{i=2}^{n} P(X_i = x_i | X_{i-1} = x_{i-1})$$

In concrete, whenever we model a language, in a bigram model, we condition the probability of a particular word occurring on that position on the previous word that we have seen. So for instance, we want to determine the maximum likely word to occur given that the previous word is "the". For this, a table is built up using a training data and all the words have an associated probability with the previous word. We can infer by looking up this table that the most likely word that follows the is going to be X. Problems arise when a word not occurring in the training data set occurs in the test set. Certain class of smoothing techniques are applied. More about smoothing techniques can be read in [6].

The language model is evaluated with a measurement called perplexity. The expression for perplexity is given by: 2^H where H represents the entropy of the model. Here the entropy is the combination of probability of D, the language sample.

$$H = -\sum_D P(D).\ logP_M(D)$$

Here, represents probability over D the new sample of sentence and represents the probability that D represents the language in the model. Using this parameter, we evaluate the different models in language modelling. Decision trees fell short of expectations reducing the perplexity factor by 4% compared to baseline trigram models. The maximum entropy models (exponential models) are marked with significant success because they managed to reduce perplexity by a factor of 39% compared to the convention bigram model.

3.2 POS tagging

The parts of speech tagging is the task that is a precursor to the task of parsing. The meaning of this phrase is that a word in the sentence is tagged or labelled with a part-of-speech. More concretely, the POS is the process of assigning a lexical class marker to each word in the sentence according to the context. The lexical class that is assigned to each word are types like noun, pronoun, adjective, verb among others. There are broadly two methods namely Rule based and stochastic. Rule based approach uses a large database of handwritten disambiguation rules considering the morpheme ordering and contextual information [8].Rule-based tagger use linguistic rules to assign the correct tags to the words in the sentence or file, e.g. verb identification rule, noun identification rule, pronoun identification rule, adjective identification rule [7]. Due to the manually written rules, rule based taggers are complex and time consuming and hence stochastic methods are preferred over rule based. Statistical methods mainly are divided into three parts namely HMM (generative model), maximum entropy and conditional random fields.

For the parts of speech tagging the most widely used algorithm is the Viterbi algorithm while considering the HMM. The Viterbi algorithm builds up on the principle of dynamic programming and probability model of the language. The maximum entropy models show 96.6% accuracy for previously unseen tags. [1] The HMM model has a baseline performance of 90% in identifying unknown tags. [2]

3.3 Parsing

The preliminary language processing culminates when the parse tree is generated. An example of parsing is the generation of parse tree showing the relation between different components of a sentence. As an example consider the sentence, John hit the ball. Here, to establish a relation between different words, we need a parse tree that does that

for us. One parse tree that does that is shown in the diagram below.

The parsing is not so simple because the language that we use has a grammar that is inherently ambiguous. This causes different parse trees that can mean different things. As an example consider the sentence, "Happy cats and dogs live on the farm." As depicted in the diagram, we can have two parse trees that have distinct meanings: one means that cats and dogs both of the species that are happy, live on the farm; the other meaning is that happy cats and ordinary dogs live on the farm. Because of this ambiguity, the meaning changes and hence parsing aims to eliminate or, at least reduce the ambiguity caused by the ambiguous grammar.

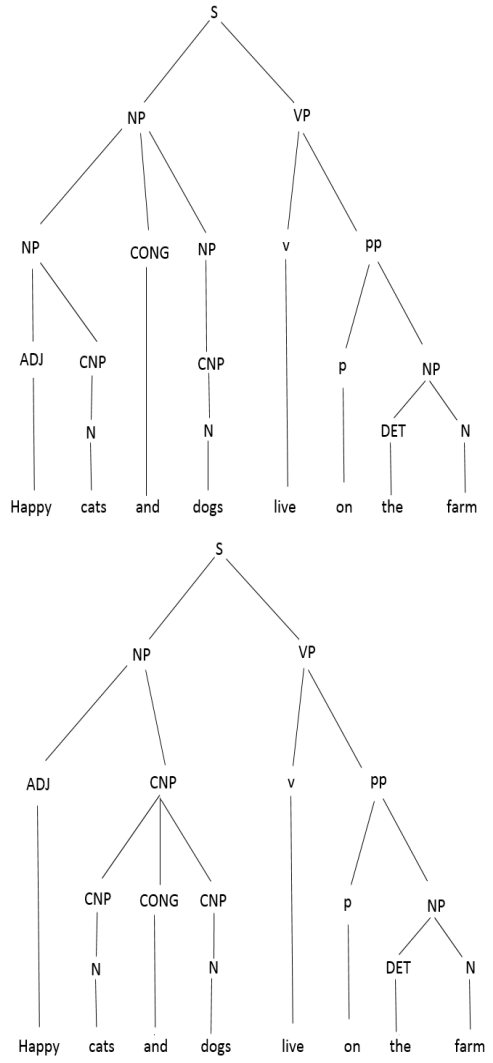

Figure 2 [4]

For the purpose of parsing, there are various methods available, however, we will delve into the one of the most recent method as proposed by Michael Collins [3]. Here, the lexicalized parser is given by a set of context free grammar rules with a probability associated with each grammar rule. These are the similarities it shares with the probabilistic context free grammar models. In addition to these probabilistic rules of grammar, lexicalized PCFG also have head associated with each rule that carries a lexicalized meaning to the upper(parent) node.

The main advantage of this extension is that the lexicalized information is retained while parsing a sentence and hence the attachment of phrases is more easily performed than what was possible with just PCFG. More formally, a lexicalized PCFG is given by a grammar that consists of non-terminals, terminals, production rules, start state.

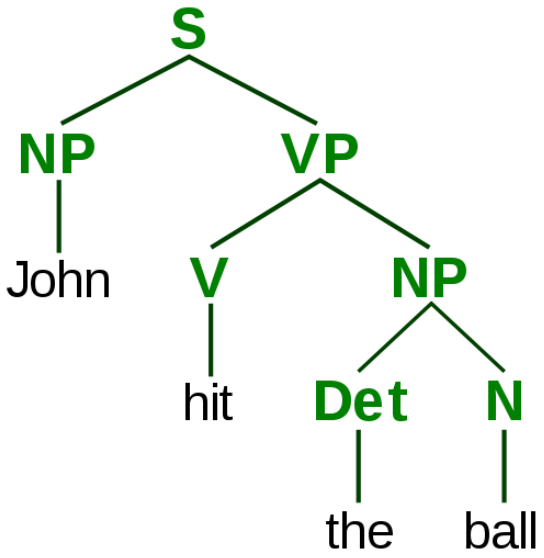

Figure 3

Each rule has a probability associated with it and a head that is served from the parent. As an example consider this instance of a lexicalized grammar rule.

S(vp)-> NP VP 0.9

This parsing model was tested on the Penn tree bank that consisted of forty thousand sentences. The measurement is based on the precision and recall values of the model. The results of the same computing with the previous models and often outperforming other models of parsing by a margin of a few percentages. The precision and recall of the traces found by the model were 93.8% and 90.1% respectively. [3]

4. REFERENCES

[1] Adwait Ratnaparkhi, A Maximum Entropy Model for Part-Of-Speech Tagging

[2] D Jurafsky, JH Martin, Speech and Language Processing.

[3] Michael Collins, Head-Driven Statistical Models for Natural Language Parsing

[4] Bill Wilson, University of New South Wales.

[5] Roni Rosenfeld, Two decades of statistical language modeling: where do we go from here?

[6] Stanley F. Chen, Joshua Goodman, An Empirical Study of Smoothing Techniques for Language Modeling. Proceedings of the 34th Annual Meeting of the ACL, June 1996

[7] Nidhi Adhvaryu, Prem Balani, Survey: Part-Of-Speech Tagging in NLP, International Journal of Research in Advent Technology (E-ISSN: 2321-9637)

[8] Dinesh Kumar, Gurpreet Singh Josan," Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications Volume 6–No.5, September 2010, pp. 1-9

[9] Manish Shrivastava and Pushpak Bhattacharyya, Hindi POS Tagger Using Naive Stemming: Harnessing Morphological Information Without Extensive Linguistic Knowledge, International Conference on NLP (ICON08), Pune, India, December, 2008

[10] PVS Avinesh, G Karthik, "Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning" in the proceedings of NLPAI Contest, 2006

[11] Antony P.J, Santhanu P Mohan, Soman K.P,"SVM Based Part of Speech Tagger for Malayalam", IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 339-341, 2010

[12] Agarwal Himashu, Amni Anirudh," Part of Speech Tagging and Chunking with Conditional Random Fields" in the proceedings of NLPAI Contest, 2006

[13] Brants, TnT – A statistical part-of-speech tagger. In Proc. of the 6th Applied NLP Conference, pp. 224-231, 2000

[14] Cutting, J. Kupiec, J. Pederson and P. Sibun, A practical partof-speech tagger. In Proc. of the 3rd Conference on Applied NLP, pp. 133-140, 1992

[15] Sumam Mary Idicula and Peter S David, A Morphological processor for Malayalam Language, South Asia Research, SAGE Publications, 2007