Chapter 3

PART OF SPEECH TAGGING TECHNOLOGIES

Parts of speech (POS) tagging means assigning words with its appropriate parts of speech in a Natural language sentence. POS tagging is very useful for language translation and information extraction. There are different approaches to the problem of assigning a part of speech (POS) tag to each word of a natural language sentence and in this section we shall discuss different part of speech tagging technologies.

3.1 Different approaches for POS tagging

There are different models for part of speech tagging. It can be classified as Supervised and Unsupervised. Both the supervised and unsupervised models can be classified as rule-based and stochastic model. The following fig 3.1 demonstrates different [30] POS tagging models.

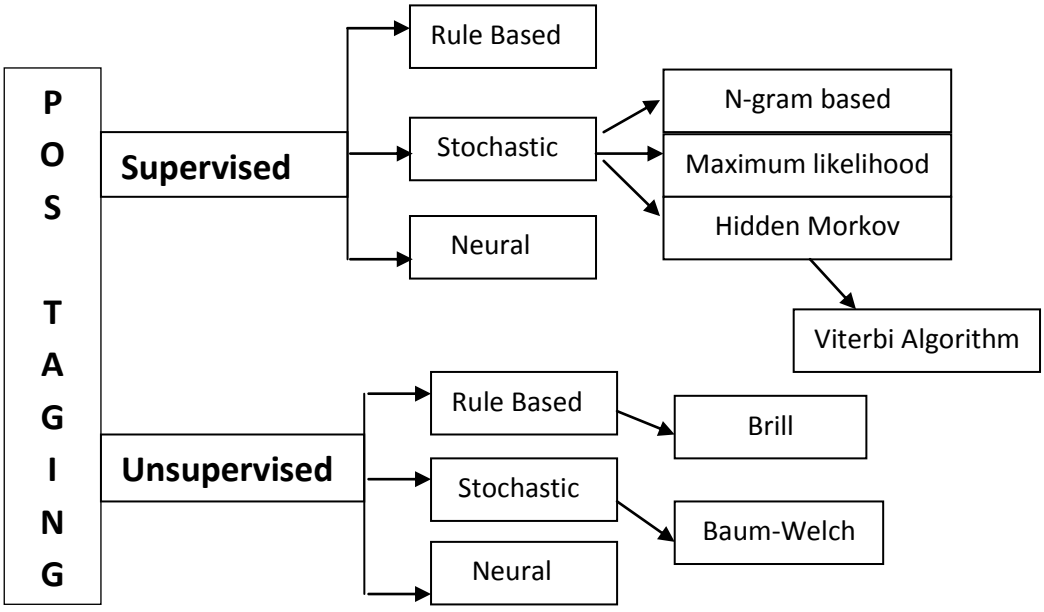

Fig 3.1 Classification of POS tagging models

3.2 Supervised and Unsupervised Part of Speech Tagging

Supervised POS Tagging: In supervised POS tagging models a pre-tagged corpora is required, which is used for training to learn about the tagset, tag frequencies word, frequencies, rule sets etc. The accuracy or performance of these models generally increases when we increase size of the corpora.

Unsupervised POS Tagging: The Unsupervised POS tagging model do not require pre-tagged corpora and they use advanced computational methods to automatically induce tagset, transformation rules etc and able to calculate the probabilistic information which is needed for stochastic taggers or contextual rules needed by rule-based systems. Unsupervised POS tagging avoid costly annotation. Despite its recent progress, the accuracy of unsupervised POS taggers still falls far behind supervised systems, and is not suitable for most applications.

3.3 Rule Based Approach

The basic principle of rule based approaches is that, the knowledge base consists of a set of linguistic generalizations, known most commonly as rules or constraints. Each rule contains the instructions for an operation to be performed, and the context describing where the rule should be applied. And these rules are responsible to provide the appropriate tags to the text. Typical rule based approaches use contextual information to assign tags to ambiguous words or unknown. These rules are often known as context frame rules. As an example, “if any word is preceded by a determiner and followed by a noun, then it is tagged as an adjective”.

In addition to contextual information, many taggers use morphological information to aid `in the disambiguation process. One such rule might be: “if an ambiguous/unknown word ends in an -ing and is preceded by a verb, label it a verb” (depending on your theory of grammar, of course). Some systems go beyond using contextual and morphological information by including rules pertaining to such factors as capitalization and punctuation. Information of this type is of greater or lesser value depending on the language being tagged.

3.4 STATISTICAL (STOCHASTIC) APPROACHES:

The term 'stochastic tagger' can refer to any number of different approaches to the problem of POS tagging. Any model, which applies probability methods, i.e. statistics, may be labelled as stochastic. In the statistic model, the tag encountered most frequently in the training set is the one assigned to an ambiguous instance of that word. This type of model or tagger is of very high accuracy but the problem with this approach is that, while it may yield a valid tag for a given word, it can also yield inadmissible sequences of tags. Another problem of this model is to improve its performance.

The alternative way to calculate the probability of a given sequence of tags occurring is word frequency approach. This type of approach is referred to as the n-gram approach, referring to the fact that the best tag for a given word is determined by the probability that it occurs with the n previous tags. Viterbi Algorithm is the most dynamic and common algorithm for implementing an n-gram approach of POS tagging. Viterbi Algorithm is a search algorithm which uses the best N Maximum Likelihood Estimates (where n represents the number of tags). If n is one (i.e. 1-gram approach), it is just the word frequency. If n is two (2-gram approach), it has a special name: bi-gram approach.

For example, if we consider the frequency of “the smile”, where “the” is determiner and “smile” is common noun, this approach is bi-gram approach. Similarly, if we consider frequency of words of order n, it is n-gram approach. The next level of complexity that can be introduced into a stochastic tagger is the one that combines the previous two approaches, using both tag sequence probabilities and word frequency measurements.

3.4.1 Hidden Markov Model

A Hidden Markov Model is a statistical Markov Model [32] and Markov chain is very useful when we compute a probability for a sequence of events. In some cases, the events we are interested in may not be directly visible in the world. For example in part of speech tagging, tags cannot be observed in the word; only words can be seen and the right tags to be inferred from the sequence In simpler Markov models the state is directly visible to the observer, so the state transition probabilities are the only parameters. But in Hidden Markov model the state is hidden to the observer and output produced with the help of those states which is visible to the observer. Each state has a probability distribution over the possible output tokens or words. Therefore, the sequence of tokens generated by a Hidden Markov Model gives some information about the sequence of states.

According to Rabiner in 1989, HMM Model has five elements and they are as follows:

1) The number of distinct states in a model is N and for Part-of-speech tagging, N is the total number of tags used by the system. Each tag consists of one state in the tagset [10].

2) In HMM, the total number of distinct output symbols are represented by (M), and the individual symbol denoted as $V = \{v_1, v_2, ..., v_M\}$.

3) The state transition probabilities can be represented by the symbol A = {aij} where probability of moving from states i to j. For part-of-speech tagging, state transition probability will be move from tag ti to tj and the probability is estimated from the corpus.

4) The observation probability which is represented by B = {bj(k)}.The probability bj(k) of having symbol k on state j.

5) The initial state distribution πi is the probability where the model will start in state i. This is the probability for POS tagging, that the sentence will start with tag ti. When HMM model performs POS tagging, our aim is to find out the most likely tag or the Viterbi algorithm can be used to find out the most likely tag sequence [10].

For finding the maximum probability HMM uses the Viterbi algorithm. The main idea of the Viterbi algorithm is that instead of iterating over all possible state sequences to choose the best state sequence, we iterate over all possible candidates of each state to get the best one for that individual state. The concatenation of the best individual states produces the best state sequence. Applied to tagging, this algorithm searches for the best tag for each word in order to find the best tag sequence.

3.4.2 Conditional Random Fields

Conditional Random Field (CRF) is a framework of probabilistic model to segment and label a sequence of data. A conditional model is a undirected model which specifies the

probabilities of possible label sequences given a particular observation sequence. The label sequence of CRF can depend on arbitrary, non-independent features of the observation sequence. The probability of a transition between labels depends on the current observation [64], and also on past and future observations [31]. The CRF model calculates the probability based on some features, which might include the suffix of the current word, the tags of previous and next words, the actual previous and next words etc. [37]

3.4.3 Maximum Entropy Model

Maximum Entropy (ME) is a very flexible method of statistical modelling. In machine learning, a maximum-entropy Markov model (MEMM), or conditional Markov model (CMM), is a graphical model for sequence labelling that combines features of hidden Markov models (HMMs) and maximum entropy (MaxEnt) models. An MEMM is a discriminative model that extends a standard maximum entropy classifier by assuming that the unknown values to be learnt are connected in a Markov chain rather than being conditionally independent of each other. MEMMs find applications in natural language processing, specifically in part-of-speech tagging[31] and information extraction.[32].The probability model for MEM is defined as:

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

where H is the set of possible word and or “histories”, and T is the set of allowable tags. π is a normalization constant, {a1, ..., ak} are the positive model parameters, {f 1 , . . , fk} are known as “features”, where fj(h, t) is in {0, 1}and each parameter aj corresponds to a feature fj.

Given a sequence of words {wl , . . . , wn} and tags {tl,..., tn} as training data, hi is defined as the history available when predicting ti. The parameters {a1, …, ak} are then chosen to maximize the likelihood of the training data p [30], using the following formula:

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)}$$

An advantage of MEMMs over than HMMs for sequence [63] tagging is that they offer increased freedom in choosing features to represent observations. Another advantage of MEMMs over HMMs and conditional random fields (CRFs) is that training can be considerably more efficient.

3.4.4 Memory Based Learning

The Memory Based Learning (MBL) Model takes tagged data as input, and produces a lexicon and memory based POS tags as output.MBL consists of two components, one is a memory based learning component, and the other is a similarity based performance component. The learning component is called memory based as it memorizes examples while training. The performance component matches the similarity of the input with the output of the learning component to produce the actual output of the system [31].The

different models described above have their own advantages and disadvantages, however, they all face one difficulty, which is to assign a tag to an unknown word which the tagger has not seen previously i.e. the word was not present in the training corpora. Different tagging models use different methods to get around this problem. The rule based taggers use certain rules to specially handle unknown or ambiguous words. But the stochastic taggers have no way to calculate the probabilities of an unknown word beforehand. So to solve the problem, the taggers of this category calculate the probability that a suffix of an unknown word occurs with a particular tag. If HMM is used, the probability that a word containing that suffix occurs with a particular tag in the given sequence is calculated. An alternate approach is to assign a set of default tags to unknown words. The default tags typically consist of the open classes, that are word classes, which freely admit new words and are readily modified by morphological processes [33], examples are Noun, Verb, Adjective, Adverb etc. The tagger then disambiguates using the probabilities that those tags occur at the end of the n-gram in question. A third approach is to calculate the probability that each tag in the tagset occurs at the end of the n-gram, and to select the path with the highest probability. This, however, is not the optimal solution if the size of the tag set is large [37].

3.5 Hybrid Part of Speech tagging

Hybrid models are basically combination of rules based and statistical models and makes new methods using strongest points from each method. It is makes use of essential feature from ML approaches and uses the rules to make it more efficient. Hybrid methods are ideally to be used to increase the accuracy of the system.

CLAWS [39] system is a good example of hybrid approach. In CLAWS1, the WORDTAG lexical analysis component has initially assigned potential tags which

were altered by rule based component IDIOMTAG. After that a stochastically disambiguator was applied (Hardie 2003). CLAWS system gives an example of hybrid approach in which both rule based system and stochastic system were developed together. Results were found to have accuracy rate of 98.5% which were better than any of the tagger.