This paper proposes new parameter estimation algorithms for tagging problems. They are variants of the perceptron algorithm, and they rely on Viterbi decoding with (voted/averaged) additive updates. Theorems about the convergence of these algorithms are discussed. Moreover, as alternatives to popular Maximum-entropy (ME) and Conditional Markov Random Fields (CRFs), the algorithms give promising experimental results on part-of-speech tagging (POS) and base noun phrase chunking.

In maximum-entropy models, the tagging task is represented through a (local) feature vector representation(phi) of history-tag pairs. “History”, a 4-tuple (2 previous tags, the array specifying the words in the input sentence and the index of the word to be tagged), is the context in which the tagging decision is made. On the basis of “local” representation phi (for history-tag pairs), “global” representation Phi (for word sequence-tag sequence pairs) is obtained by summing phi over all history-tag pairs in word sequence-tag sequence pairs. Typically, if local features are indicator functions, then global features will be “counts”.

In the proposed model, the “score” of a tagged sequence for parameter estimation is expressed as the dot product of the parameter vector and the global feature Phi. It is different from conventional maximum-entropy method by omitting the local normalization term. The parameter estimation method takes a set of n tagged sentences as input and learned parameter vector (alpha) as output. During T training loops for n sentences, Viterbi algorithm is applied to find the output of the model on the ith training example with current parameter settings (z, the tag sequence maximizing the “score” among all feasible tag sequence for this sentence), and update the parameter vector when z is different from the true tag sequence (t). Intuitively, parameters are increased when corresponding features are “missed”, decreased when “incorrect”, and unchanged when “correct”. Furthermore, averaging parameters over iterations and training examples for all entries refines the algorithm and brings better performances.

In the section (3, 5) for theories justifying the algorithm, the author first provides a general algorithm for which abovementioned algorithms are special cases by setting specific training examples, GEN and feature representations. Then, theorems bounding the number of errors on training examples are discussed. When the training set is “separable”, then the algorithm will converge to parameters values with 0 training error. If training set is just close to being separable, then the algorithm is still able to just make a small number of mistakes. It is worth mentioning that the number of mistakes is not dependent of the number of tag sequence candidates (can be exponential), but dependent only on the separation of the training set. Moreover, for generalization on new test data, there are theoretical results implying that if the algorithm makes relatively small error, then it is possible to generalize well on test data. During the discussion, the concept of “voted perception” is introduced, to which the averaged algorithm is an approximation.

Experiments are conducted on different data sets for POS and base noun-phrase recognition for both ME method and the proposed perceptron algorithm. Variants of each algorithm families are explored with various parameter settings. On the development data set, for the perceptron algorithm, averaging improves results significantly and help stabilize the tagger by including all features. In contrast, ME suffers with all features included. Moreover, on the test data set, the perceptron algorithm gives better performances on both tasks (11.9%, 5.1 % relative error reduction) compared to ME.