Working of a POS Tagger & the Entropy Model

Introduction:

The author proposes a Maximum Entropy Statistical Model that comprises training from a corpus annotated with the Parts of Speech Tags and using the same on the previously unseen text. The author claims that the model clocks an accuracy rate of 96.6%. Also, the author discusses the features to handle difficult tagging decisions as well as to address the corpus consistency problems.

Working of a POS Tagger:

The Probability Model:

- The author defines a probability distribution over a sample space of H*T where H is the set of words and tags & T is the set of allowed tags where each of the tuples encountered (h,t) h belongs to H & t belongs to T.
- These define the presence or absence of the features which in turn define the inclusion of feature parameters in the distribution.
- These parameters are defined by the maximum likelihood estimation over the joint distribution of h and t for all the features. These features are defined in such a way that it can define the co-occurrence of a tag 't' concerning the word 'h'.
- The author then goes on to translate this probability model into an entropy-based model by defining the entropy of the probability distribution as the entropy of the joint probability of h and t of the expected values of the model and observed features.
- If the probability model satisfies the following conditions:
  - ➔ K (Constant Value) constraints
  - ➔ Maximizes the computed entropy over the distributions

  then it uniquely maximizes the distribution over the defined Probability Model.

Features:

- The author defines the join probability of history H & tags T based on those parameters whose features are active.
- If those features exist, the model parameter contributes towards the joint probability.
- This feature can activate on any word or tag in the given history and helps predict the correct spelling of the word as well as the previous two tags.
- For tagging unknown words, it is assumed that these are similar to the unknowns in the test data

Testing

- The Test Corpus is tagged using one sentence at a time
- This involves searching for the candidate tag sequence for the sentence and the one with the highest probability is selected as the answer.
- The tagging for the Test Corpus is done via a search algorithm called "Beam Search" that uses the conditional tag probability.
- The above-mentioned algorithm maintains a Tag Dictionary for each word and consists of a list of all the tags associated with that word in the training set.
- The search procedure uses the Tag Dictionary to tag the words in the test corpus using the data in the dictionary. For the unknown words encountered in the test corpus, all the tags are used to associate with that unknown word.