Part 2 : Part-of-Speech Tagging using Hidden Markov Model (HMM)

Part of Speech Tagging (POS) is a process of tagging sentences with part of speech such as nouns, verbs, adjectives and adverbs, etc.

Identifying part of speech tags is much more complicated than simply mapping words to their part of speech tags. It is quite possible for a single word to have a different part of speech tag in different sentences based on different contexts.

POS-tagging algorithms fall into two distinctive groups:

1. **Rule-Based POS Tagger:** For the words having ambiguous meaning, rule-based approach on the basis of contextual information is applied. It is done so by checking or analyzing the meaning of the preceding or the following word. Information is analyzed from the surrounding of the word or within itself. Therefore words are tagged by the grammatical rules of a particular language such as capitalization and punctuation. For example, if the preceding word is an article, then the word in question must be a noun. This information is coded in the form of rules. e.g., Brill's tagger.

2. **Stochastic POS Tagger:** Different approaches such as frequency or probability are applied under this method. If a word is mostly tagged with a particular tag in training set then in the test sentence it is given that particular tag. The word tag is dependent not only on its own tag but also on the previous tag. This method is not always accurate. Another way is to calculate the probability of occurrence of a specific tag in a sentence. Thus the final tag is calculated by checking the highest probability of a word with a particular tag. This is sometimes 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. This approach makes much more sense than the one defined before, because it considers the tags for individual words based on context.

The next level of complexity that can be introduced into a stochastic tagger combines the previous two approaches, using both tag sequence probabilities and word frequency measurements. Tagging Problems can also be modeled using HMM. It treats input tokens to be observable sequence while tags are considered as hidden states and goal is to determine the hidden state sequence.

For example $x = x_1, x_2, x_3, x_4, ...x_n$ where x is a sequence of tokens while $y = y_1, y_2, y_3, y_4, ...y_n$ is the hidden sequence. HMM uses join distribution which is $P(x, y)$ where x is the input sequence/ token sequence and y is tag sequence. Tag Sequence for x will be argmax $y_1 ....y_n \; p(x_1, x_2, x_3, x_4, ...x_n, y_1, y_2, y_3, y_4, ...y_n)$ We have categorized tags from the text, but stats of such tags are vital. So the next part is counting these tags for statistical study.

Task is to analyse the given code and report the findings corresponding to each block of the code.

https://drive.google.com/open?id=1tbCfWb99okwMn3I0RxIKGCOKUAjST0PX