Typos correction is statistical solved by HMM with Viterbi Algorithm, which is based correct letters and actual letters. In this case, correct letters are states and actual letters are observations. These elements are used in HMM for providing the states sequence from the input observation sequence.

The data used for training is a text document pre-processed. In fact, the data includes lower case letters and space character and it is stored in the form of array. An entity is an array of two letters being correct and actual letter. The dataset has the training and testing set with 10% error and 20% error.

HMM is an approach of sequential pattern recognition and described as a matrix of states with the probability of state transition and observation. With N distinct states, qt is the state of the time instance t. In the first order Markov chain, the state transition probability depends only on its previous state. Therefore, P(qt =i| qt-1, qt-2 ....) = P(qt | qt-1) and a state transition ai,j = P(qt =j| qt-1 = i). The the second order Markov, the state transition probability depends on the two previous states ak,i,j = P(qt =j| qt-1 = i, qt-2 = k ).

These probabilities are represented by a transition probability matrix A and initial state distribution vector π.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | A = | a11 | a12 | ... | a1N | | a21 | ... | ... | ... | | ... | ... | ... | ... | | aN1 | aN1 | ... | aNN | | |  |  | | --- | --- | | π = | π1 | | π2 | | ... | | πN | |

In addition, HMM has observation probability bj (ot) = P(ot | qt = j). As a result, the model λ = (A, B, π).

The task is predicting correct letters from actual letters. In this case, with the observation sequence o = (o1 , o2, ... oT ) and λ = (A, B, π), the task becomes how to choose the corresponding state sequence q = (q1, q2, ... qt) in an optimal sense.

Based on dynamic programming, Viterbi algorithm provides the result sequence by using the maximization δt(i) = max P (q1 q2 ...qt-1 ,qt = i ,o1 o2 ...ot| λ).

In the case of typos correction, with training set and test set 10% of error, the number of errors that the model first HMM correct is 0.9443989 and 0.951366.