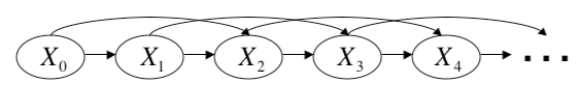
**Assignment #2**

1. *Generate sentences using this model by implementing the prior-sample routine on this network.*



**Fig. 1 Second order Markov Model.**

In order to generate sentences, we will use a second order Markov Model, seen in Fig. 1. This model states that the probability of is dependent solely on and , or in our particular example, the next word would be dependent solely on the preceding two words.

Therefore, based on the trigram data file, which contains the second order conditional probabilities (ie. ), we can obtain a list of possible next words. From this list of possible next words, the probability that a certain word will occur, , is computed using equation 1.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

The probabilities for each next possible word are then fitted to a Normal distribution (considering only the left half of the probability distribution curve) and a random number is sampled from this distribution.

Below are a set of examples of sentences generated that have been generated using the model described above:

<s> I have no idea that he is not the case . </s>

<s> The Admiral ' s character , and I have been a great deal to be the same . </s>

<s> I have not been able to say , that we should not have been so very well . </s>

<s> " Oh ! </s>

<s> I would have been in a very pretty girl , with the utmost ; and , with the most perfect alacrity he welcomed the relationship , alluded to my mother , and the first , and I do . </s>

<s> Do not let me be a very good sort -- which was , and I have not seen a great deal to say so , I hope you will be the same . </s>

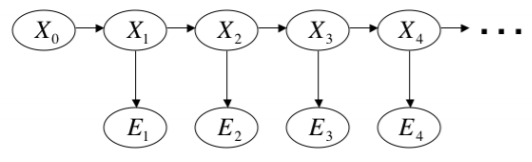
<s> I have no idea of his wife . </s>

<s> Her mind was not more ; but I was very well . </s>

<s> I should like it , I have not seen Mr . Weston ' s , I do not know . </s>

<s> I am sure I should like it ? </s>

1. *Write a program to correct noisy (incorrect) input sentences using this model by implementing the Viterbi algorithm for most likely sequence computation.*



**Fig. 2 Basic Hidden Markov model given observations .**

The procedure for correcting the noisy sentence inputs to the program uses a Hidden Markov Model, as seen in Fig. 2. Using the bigram distributions relating the current word that we are loking for and the observation word, we can determine the probability of each of the possible words using the edit distance (or Levenshtein distance).

The Levenshstein edit distance used in the calculation of was obtained from <https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance#Python> which was the dynamic programming implementation in Python of the number of corrections needed between the two strings in order for them to be the same.

Using this distance we obtain which we use together with to determine the probability of the current word. The Viterbi algorithm in the equation below was used to determine the correct sentence given a noisy input string .

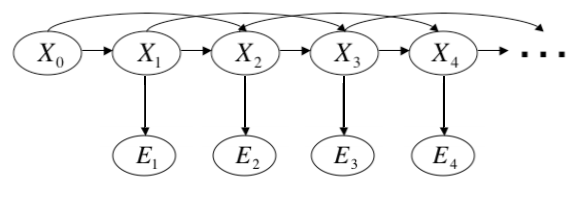
Below are the noisy sentences that have been corrected using the procedure described above:

|  |  |
| --- | --- |
| **Noisy Input** | **Corrected Sentence** |
| <s> I think hat twelve thousand pounds | <s> I think that twelve thousand pounds  \*\*\*Note: I was obtaining this sentence before adding a slight correction that may have made the resulting other sentences wrong.  <s> I think that there should you |
| <s> she haf heard them | <s> she had heard them |
| <s> She was ulreedy quit live | <s> She was already quite like |
| <s> John Knightly wasn’t hard at work | <s> John Knightley was hard at work |
| <s> he said nit word by | <s> he said it would be |

Annotated output file for each sentence can be found in files *A2\_sentenceOutput<#>.txt*, where # is 1-5.

\*\*\*This part can be disregarded as I tried it once for fun but the results came up drastically different for some of the sentences. It was not part of the assignment description though.

This procedure can be improved further by using instead of using instead of simply as the distribution for the . This would allow the prediction of the correct word to be dependent on the last two words rather than just on the previous one.



**Fig. 3**

Below are the noisy sentences corrected using

|  |  |
| --- | --- |
| **Noisy Input** | **Corrected Sentence** |
| <s> I think hat twelve thousand pounds | <s> I think it twelve hundred ; |
| <s> she haf heard them | <s> she was heard three |
| <s> She was ulreedy quit live | <s> She was already quite alive |
| <s> John Knightly wasn’t hard at work | <s> John Knightley was hard at work |
| <s> he said nit word by | <s> he had not good company |