COMP 550 Natural Language Processing

Assignment #2 - Christopher Zheng - 260760794
christopher.zheng@mail.mcgill.ca

October 15, 2019

1. Q1

   (a) Claim: Viterbi does produce the global optimum.

   Prove by induction.

   Base case: step number = 1.

   When there is only one step between two ends, Viterbi can select the path with the max probability according to the Viterbi definition. Thus, Viterbi produces a local maximum at $n = 1$

   Induction hypothesis: assume this claim is true for all n previous steps. Inductive step: step number = $n + 1$.

   As we already have the best path of all n previous steps and Viterbi selects the max for the step of $n + 1$, by the Markov property of HMM (we can make predictions for the future of the chain based on only its present state just as well as we could knowing the chain's full history), the selected path is globally optimal for the previous $n + 1$ steps.

   Therefore, Viterbi can produce the global optimum. □

   (b) Claim: the two given expressions are equivalent.

   The cross entropy loss incurred each timestamp t can be expressed as $p_t^i = P(Y_t|X)$ and this suggests that we have marginalized all Ys that are not for t. That is, given t and $Y_t = j$, $\forall y$ possible sequences, $P(Y_t = j|X) = \sum_{y\, where\, Y_t=j} P(Y = y|X)$. If we take the logarithm and sum up the cross entropy losses over all timestamps taken by a sample, we may get the result of the addition as $-\sum_t log(P(Y_t|X))$. Then, we add up all samples' cross entropy loss and may obtain $-\sum_i \sum_t log(P(Y_t|X))$ where i's denote the samples. The meaning of this sum is to compute the loss of each timestamp in the sequence with cross-entropy and to sum them up as the loss for the overall sequence.

   This sum is equivalent to the negative log likelihood loss for linear-chain conditional random fields $-\sum_i logP(Y^i|X^i)$ which is to compute the loss of a sequence as a whole entity all the time. □

2. Q2

(a) 1. What are some advantages of modelling French grammar with a CFG?

A: Since the structure looks really like pattern matching, it is relatively easy and accurate to parse for parsers and compilers. Besides, this representation is space and time efficient.

(b) 2. What are some disadvantages of modelling French grammar with a CFG?

A: The conditions of all the branches are not very readable if the user has no linguistics or functional programming background. Plus, the breaking down branches are too specific to be encapsulated and applied to other scenarios. Finally, modifications to such a structure may lead to re-building the whole parse tree.

(c) 3. What are some aspects of French grammar that your CFG does not handle?

A: Some proper nouns that come with articles except for le Canada may crash the system. (e.g. le Japon, and la Chine)

3. Q3 Decipherment Report

In this work, we experiment on three ciphers with the Hidden Markov Model (HMM) as well as multiple techniques, attempting to improve the overall decipherment performance. We examine the accuracies (shown below) on the three ciphers, and compare the methods that we utilize.

1) As a general approach, we take advantage of a standard Hidden Markov Model. The result is surprising to some extent, as it achieves a higher accuracy on Cipher 3, which is the most complex in terms of its structure. We speculate the poor performance on simpler ciphers may be due to the limited size of the training corpus. 2) To improve the standard HMM, we add the Laplace smoothing and assess the accuracies. This methods brings great accuracy boosts for both Cipher 1 (which achieves close to perfect accuracy) and Cipher 2, but witnesses no increase in Cipher 3. This result is relatively expected and may be explained by the fact that Laplace smoothing, which tends to increase the probabilities of unforeseen events, is more effective in handling brand new patterns for simple ciphers which encode sentences in a more straight-forward approach. 3) Attempting to further enhance the accuracies based on the standard HMM, we include a new corpus, which consists of two short English novels, to supplement our experiments. We witness a large boost in accuracy for first two ciphers and almost no change on the last cipher. 4) We incorporate the improved model along with the new corpus and notice a small increase in accuracy compared to 2) for the first two ciphers but not the last one.

To summarize, the Hidden Markov Model can achieve decent results in decipherment with proper implementation, given a training corpus of a considerable size. We note that the model cannot deal with Cipher 3 because Cipher 3 breaks the Markov Property that the future state is associated with only the current state.

	Cipher 1	Cipher 2	Cipher 3
Standard HMM	0.099	0.150	0.213
HMM + Laplace	0.977	0.831	0.213
New corpus	0.684	0.607	0.219
New corpus + Laplace	0.982	0.871	0.211