Assignment 2 Report

Aditya Yele
11447727

I. Viterbi Implementation:
   - Transpose emission scores which are NxL to make them LxN and transpose transition scores matrix to make y' as columns
   - Create ans and path matrix both of size LxN.The ans matrix will store scores at each iteration in the column while the path will store the index of the row of the previous column which contributed to the max score
   - Add the emission scores and the start scores and store them in the first column of the ans
   - Iterate over the columns from second column to the end
     1. Add the transpose scores of the previous column to the transition matrix which was transposed previously. This will give us the sum for all the transitions from a tag and their scores in the cell in the previous column . i.e the term after max in equation 5.
     2. Calculate its max along the columns of the resultant matrix of the above step and add it to the emission scores of the current column
     3. Get the index of each column for every row for the resultant matrix and put it in the path matrix.
   - Add end scores to the last column of the matrix ad find the maximum score
   - Retrace the path from the path matrix by going to the index of the max score and taking the value in that place and going to the previous column and to the row of that value and adding it to the list . Go on doing this till the first column.
   - Return the list of the indices and the max score.

II. Description of added features:
   Features added
   - IS_HANDLE: Since the data is twitter data the handle names start with @will mostly be nouns so so they can be classified as such
   - IS_URLI: URLs can the classified together in to one category
   - IS_LY: Many adverbs end with ly so it can group them together
   - IS_ER|IS_EST|IS_ABLE|IS_FUL: Many adjectives end with "er"-better, "est"-strongest,"able"-capable,"ful"-beautiful
   - IS_NEG|IS_IM|IS_ANTI: Words with a negative connotation like impossible, antithesis, unable, disable start with these prefixes
   - IS_EXCL|IS_Q|IS_FULLSTOP|HAS_COLON|HAS_APOS|HAS_QUOTE|HAS_COMMA: These represent a punctuation mark like a question mark, full stop, Exclamation mark, apostrophe, quote, comma. These punctuation marks tend to have certain specific type of words before or after them like the , can mostly be a

list of nouns or the : is used after a name of a person in prder to state he/she said something. “ comes at the beginning of a sentence and after the end. Also taking them together decreases the accuracy on both memm and crf.

- IS_TAG: to handle the html less than, greater that and values.
- HAS_HASH: Twitter uses hashtags and they are used mostly for nouns
- IS_SLANG: Twitter data contains many slangs which are nouns or verbs
- HAS_HYPHEN: Used to join two nouns
- IS_ED: Words ending in “ed” are used to describe verbs in past tense like happened, created
- IS_EN|IS_RE: Verbs usually are prefixed with “en” and “re” like in words like encourage and reply at times can also be nouns
- IS_ING: Words ending in “ing” are usually verbs like singing, dancing.
- IS_RT: RT is retweet in twitter will be followed mostly by noun indicating handle
- COUNT_2|COUNT_1: Words with length 2 or one are prepositions like is, as, a.

III. Comparison of your features against the basic features

Token Accuracy with basic features using memm is 84.38
Token Accuracy with basic features using crf is 84.29

The comparisons are made by first finding the accuracy of the basic model and then finding the the accuracy by adding individual features only.

In the Memm table are the top 8 best performing features.
For MEMM we have IS_LY which describes the adverbs as the top most.
It increases accuracy to 84.95.
Prominent features in MEMM are ones like symbols like exclamation mark and length of word.

No	Feature	MEMM Accuracy
1	IS_LY	84.95
2	IS_HANDLE	84.76
3	IS_CAPITALIZED	84.72
4	COUNT_2	84.72
5	IS_EST	84.53
6	IS_EXCL	84.53
7	HAS_HASH	84.53
8	IS_ER	84.48

Below are the top performers for CRF the top performering features in MEMM also present in the CRF but revers is not true. Word related features are more prominent in CRF.

No	Feature	CRF Accuracy
1	IS_ING	84.95
2	IS_HANDLE	84.67
3	IS_EXCL	84.62
4	IS_ER	84.57
5	IS_LY	84.53
6	IS_CAPITALIZED	84.53
7	IS_EST	84.48
8	HAS_HASH	84.48

IV. Comparison of MEMM and CRFs

Below is the table representing the accuracies of MEMM and CRF more each feature is added. Unlike the above the table below compounds on all the features above it

No	Feature	Dev Accuracy MEMM	Dev Accuracy CRF	Effect on accuracy Increase(I) Decrease(D)
1.	IS_HANDLE	84.76	84.67	I
2.	IS_URL	84.81	84.86	I
3.	IS_LY	85.24	85.00	I
4.	IS_ER	85.47	85.33	I
5	IS_CAPITALIZED	85.38	84.76	I
6.	IS_OUS	85.43	85.38	D
8.	IS_EST	85.57	85.90	I
9.	IS_EXCL	85.61	85.43	I

10.	IS_Q	85.61	85.47	~Same
11.	HAS_COMMA	85.57	85.66	DI
12.	HAS_APOS	85.43	85.90	DI
13.	HAS_QUOTE	85.47	84.57	D
14.	IS_ABLE	85.43	85.90	DI
15.	IS_FUL	85.47	85.52	I
16.	IS_EN	85.43	85.71	DI
17.	IS_NEG	85.38	85.14	D
18.	HAS_HYPHEN	85.61	85.52	I
19	HAS_HASH	85.71	85.66	I
20	IS_RE	85.66	85.99	I
21	IS_ING	85.52	86.56	I
22	COUNT_2	86.41	86.09	ID

The effects that the combined features have on both accuracies is different than the basic model. For example OUS and EST together increase accuracy while individually decrease.
Also for MEMMs the combination of symbols like punctuation marks and length of words increases accuracy while decreases the accuracy for CRF.
For CRF strong text based features seem to work more on general language features like adjectives which mostly end with “est”, “able”,”ful”.
Also Twitter specific features like hash and @ increase the accuracy for both types.
The highest overall accuracy achieved on CRF was 86.56 while on MEMM it is 86.42.
Thus overall CRF performs better than MEMM.

Comparison with Basic features.
Basic features are made on the structure of the text like if they are numeric or alpha numeric or upper or lower case.
My features are based on the estimating the type of text like the clustering together ly assuming they will belong to one category. Also they into account the symbols that will be present and the words around them may belong to a specific category.

Sentences

1. MEMM
   For MEMM sentences which consists of symbol and length of two words seem to dominate accuracy.
   Hey! We as a team @loner!!!
   #Teamwork

   This sentence has more symbols and two or one length words.

   Ah! What a surprise!! #fun
   Oh? It is as it is!

   Fox:”Nobody can catch me!!!”

   2. CRF
   For more characteristic features of parts of speech will work along with usual twitter specific symbols

   A beautiful day and a lovely evening. Let us quickly have dinner.
   Here we have words like lovely quickly and beautiful which are adverbs and adjectives which give more accuracy on CRF

   RT: :Strongest and the quickest shall win. @Darwin