2 Most frequent tag baseline

2.b

F1=0.8

3 HMM tagger

3.b

In order to make the actual runtime better, we chose to use pruning on some of the words. We implement this idea by using ‘max size’ parameter, which eliminates some tags for specific words: at any timepoint, for any specific index ‘i’ in the sentence, there are only ‘max size’ sets of tags (which has the best probability till index i). We arbitrarily set the max size to 42.

Optimal lambdas:

$$\lambda_1 = 0.201$$

$$\lambda_2 = 0.30100000000000005$$

$$\lambda_3 = 0.4979999999999999$$

3.c

F1=0.8366557137641476

3.d

Let’s assume we have a vocabulary of two words: ‘object’, ‘it’.

The sentence is:

$$START\ Object\ it\ STOP$$

The emission probabilities are:

$$P(w_i = object|t_i = V) = 1$$
$$P(w_i = object|t_i = N) = 0.5$$
$$P(w_i = it|t_i = N) = 0.5$$

Because ‘object’ can be either noun or verb and ‘it’ is always a noun.

The transition probabilities are:

$$P(t_i = V|t_{i-1} = N) = 0.75$$
$$P(t_i = N|t_{i-1} = V) = 0.75$$
$$P(t_i = N|t_{i-1} = N) = 0.25$$
$$P(t_i = V|t_{i-1} = V) = 0.25$$
$$P(t_i = V|t_{i-1} = O) = 0.3$$
$$P(t_i = N|t_{i-1} = O) = 0.7$$

Greedy inference algorithm would choose at each step the tag that maximizes the multiplication of emission and transition.

$$P(object|N) \cdot P(N|O) \cdot P(it|N) \cdot P(N|N) =$$

$$0.5 \cdot 0.7 \cdot 0.5 \cdot 0.25 =$$

$$0.04375$$

Yet, the maximal probability for the whole sentence is:

$$P(object|V) \cdot P(V|O) \cdot P(it|N) \cdot P(N|V) =$$

$$0.3 \cdot 1 \cdot 0.5 \cdot 0.75 =$$

$$0.1125$$

4 Maximum Entropy Markov Model (MEMM) tagger

4.d

We used both caching, on possible last two tags, and pruning the cache to hold up to 'max size' most probable pairs.

Moreover we avoided unnecessary features which didn't increased any of the P/R/F1 values but made the program much slower.

4.e

Greedy F1=0.84

Viterbi F1=0.74

4.f

We chose the MEMM Viterbi model.

Model fails mostly on the LOC tag.

It has 2 problems:

1) It assigns the tag LOC to the next word

(for example at the sentence London 1996-08-30

The word "London " is taged with 'O' and the word "1996-08-30" is taged with LOC)

2) It confuses between ORG and LOC: the word Scotland is taged as ORG.

5 BiLSTM tagger

5.b.i

If we did not use masking, the gradients from the padded input would affect the learning parameters, because they would flow through the hidden state.

To avoid the affect of the padded labels on the loss, and therefore on their gradients, we use the masks.