Information Extraction: Witten part

Q2

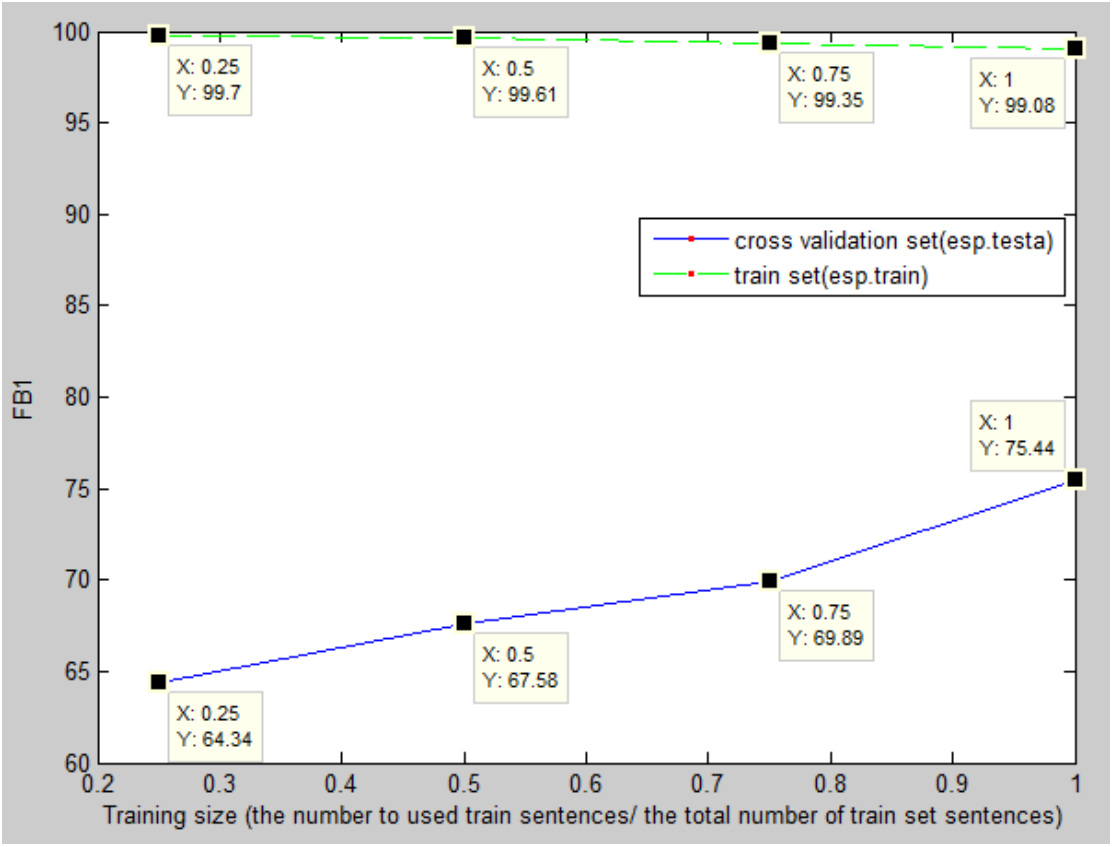

Figure 1: learning curve of the validation set (esp.testa) and train set(esp.train)

Conclusion:

The general trend is, as the size of the training set increases, the cross-validation score (FB1) of the validation set (esp.testa) increases, the training score (FB1) of the train set (esp.train) decreases.

According to the trend of the learning curve, we can continue to increase the performance of the classifier on the validation set by increasing the size of the training set.

Q3

Resubstitution performance is the training performance of a model that uses its training data to evaluate its performance, it describes how well the model fits its training data.

The green line on Figure 1 describes the resubstitution performance. As the picture shows, the FB1 of the train set decrease as the size of the training set increase, but still very high (99.08) when I use off all the training sentences. The difference of FB1 between training set and cross-validation set is very large, with a FB1 of 71.13 as the cross-validation evaluation score when I use all of the training data. This shows that the classifier doesn't generalize well and it is overfitting (with high variance). If the classifier is suffering from high variance (overfitting), more data will probably help the classifier be more generalization and thus increase the performance on test set. According to the growing trend of the cross-validation score in the learning curve (Figure 1), the FB1 will continue the increasing trend giving more training data. Thus, the classifier is overfitting.

C value can change the hyper-parameter for the CRFs. The learning algorithms can trade the balance between overfitting and underfitting with this option. If c >1, the learning algorithm tends to overfitting, if c < 1, the learning algorithm tends to underfitting.

We can judge whether the classifier is overfitting or underfitting base on its resubstitution performance and its performance on unseen validation set in the learning curve use the method I have mention above. Then we can use **c value** to adjust the classifier to help it get a better performance by reach a balance, which means if the classifier is overfitting, we can set the c value < 1 to make it underfitting during the training process to balance the overfitting, vice versa.

Q4

The hidden states (t) of my HMM are the POS tags, such as noun, adjective, pronoun, numeral, particle, verb and other.

The observations (w) of my HMM are the words in the Finnish text.

The emission probabilities, which also called observation likelihoods, each expressing the probability of an observation w_i being generated from a hidden state t_i , which can be denoted as P (w_i | t_i).

No. Instead, I would expect a larger transition probability from adjective to a noun in Finnish than in Portuguese.

Tag transition probability p (t_i | t_{i-1}) represents the probability of a tag t_i given the previous tag t_{i-1}.

In Finnish, adjectives that define a noun tend to occur before the noun, which means the transition probability from adjective to noun P (noun | adjective) in Finnish is high. While in Portuguese, adjective that define a noun tend to occur behind the noun, which means the transition probability from noun to adjective P (adjective| noun) is high, thus P (noun | adjective) ought to be low.

The transition probability is computed by the ratio of counts as follow:

p (t_i | t_{i-1}) = $\frac{C(t_{i-1},t_i)}{C(t_{i-1})}$