Assignment 4

201014560 & 057931354

Remark on Running Environment

We ran on windows on python 3.6.4. We had to make few modifications on data_util.py (add "rb" and "wb" when opening files) and and in rnn.py (replace raw_input() with input()). However, in Nova we left the original code untouched to allow it to run on python 2.7.

Question 3

(c) i – the maximum entropy value we can get is 1.0 for -SUM(i) y(t)[i] * log(y(t)[i]) where i is an index of the class (PER, LOC, etc.) and t is the time step (or a token of a given sentence). It can happen when p(i) = y(t)[i] * logy(t)[i] is 1 / n_classes, and when this is the case we get the hightest uncertainty level (y(t)[i] are completly randomized).

(c) ii – In the enropy graph we can see clearly that the value decreases as time step increases. It makes sence sence since as we trains the inpu we wish to decrease the uncertainty (randomization) level and be able to predict well. In other words we wish to be able to distinguish clearly which element in the prediction vector is the maximum (Max(y(t)[i]) and then, with high certainty, map it to the corresponding class.
The average loss graph (the J function in page 4) is very similar to the entropy graph (though a bit more noisy) Intuitivly it makes sence since as the prediction vector y'(t) becomes more distinguished (significant), we expect y(i) * log(y'(t)[i] to be higher (y(i) is the one hot vector), hence -y(i) * log(y'(t)[i] to be smaller.

In the prediction-logits histogram graph we can see that in the back, the ealier epochs or time steps, the graphs are narrow and higher and as we move to the front, the graphs become wider and lower. It makes sence since these graphs capture, at the end of the day, the percentage of the classes in the training data (the bins somehow represent the different classes). Being wider and lower tells us that the proportion between the number of classes in the training data (namely, for each class, how many instance there are in the data training) are not extreme, and at any event, that these prortion are stabilized as we move along the time. As these proportions are stabilized we expect to see the higher level of certainty or lower entropy. At the end of the day, the percentage of the different classes (for example, # ORG / total number of tokens) needs to be captured within the entropy.

(d) first we will identify the weakness of the model. We will observe few examples:

Most of	*the*	*Marines are*	*on*	*three ships in*					*the*	*Tarawa Amphibious Readiness Group*				
y:*	*O*	*O*	*O*	*MISC*	*O*	*O*	*O*	*O*	*O*	*O*	*ORG*	*ORG*	*ORG*	*ORG*	*O*
y':	*O*	*O*	*O*	*MISC*	*O*	*O*	*O*	*O*	*O*	*O*	*LOC*	*LOC*	*ORG*	*ORG*	*O*
p :	*1.00*	*1.00*	*1.00*	*0.58*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*0.90*	*0.6*	*0.65*	*0.48*	*1.00*

Here we can see that the model (y') incorrectly identified *"Tarawa Amphibious"* as LOC and not as ORG. In general, it is may be difficult to distinguish between organization and location especially in this case in which syntactically both are correct (prior to them we have the expression "in the" which can apply to both classes). In addition, it might be rare to see ORG of four words. In terms of predictions can see that the prediction of "Tarawa" as LOC is high and equal to 0.9 however the prediction

decreases to 0.6 for the next word "Amphibious" (namely, we believe "Amphibious" is LOC with probability of 0.6 which is low).

The expeditionary force would include nearly 1,000 Air Force personnel in ground and..
y: O O O O O O O ORG ORG O*
y': O O O O O O MISC MISC MISC O O O O O
p : 1.00 1.00 1.00 1.00 1.00 1.00 0.80 0.93 0.66 1.00 1.00 1.00

Here the model treated the number 1,000 as a MISC (MISC are capaital letter nouns (somehow less concrete but not necessarily) which do not fall with one of the categories of LOC, ORG or PER such as Bangladeshi or even the expression "1,000 Lakes Rally").
It is clear the 1,000, in this example above, is, unlike "1,000 Lakes Rally", not part of "Air Force".

A spokesman said he could substantially confirm a report in the news weekly Der Spiegel ,
y: O O O O O O O O O O O O O ORG ORG O*
y': O O O O O O O O O O O O O PER PER
p 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.0 1.0 1.0 1.0 1.0 1.0 0.94 0.98

Here we see the typical problem of words or, in fact expressions, being ambivalent. The model treats "Der Spiegel" as a person (PER) rather than a newspapaer (ORG).

Late on Friday Liege police said in a statement that on Thursday , Rachel Legeard , 18 , and Severine Potty , 19 , had gone shopping to the eastern town of Liege on Thursday , where Legeard 's wallet had been stolen .

This is example I believe shows the weakness and the strength of the model – the first occurrence of Liege was was wrongly labeled as ORG but the second occurrence was labeled correctly as LOC. Hardly need to mention that Liege in both occurrences refers to the city Liege.