NLP Homework 5 – Itay Itzhak, 305685877, Eytan Chamovitz, 203486550

Question 1

(a) Answer:

i. Example Sentences:

- While shopping in Amazon, one should be aware that not all vendors are legitimate. (Amazon could be a Company, a River, a geographical area, an ancient race of mythical warrior women, etc.)
- Dr. Fischer tried hard to work on a cure. (Dr. Fischer being a name, but can also be a company).

ii. Answer

- This is because a single word can have many different contexts, which affect the word's potential meanings and references.

iii. Example features:

- Preceding and following words (e.g. company denominations Inc., ltd., or titles Dr., Prof., Ms., or other)
- Capitalization of non-first word (because then it is a 'proper' noun).

(b) Answer:

i. Dimensions:

- $\vec{x}^{(k)}, k = [t-w, \dots, t+w]$ dimensions: $1 \times V$ they did not asked to write this, (and I think it's $1 \times (2w+1)V$ anyways)
- $\vec{e}^{(t)}$ dimensions: $1 \times (2w+1)D$
- W dimensions: $(2w+1)D \times H$
- U dimensions: $H \times C$

ii. Complexity

At each Iteration the calculation of $e^{(t)}$ takes $O(w \cdot D)$ operations and $h^{(t)}$ takes $O(H \cdot D \cdot w)$. For the calculation of $\hat{y}^{(t)}$ $O(H \cdot C + C)$. CE calulcation is less since it's $O(C)$,

So over since we do one forward pass across all words of a sentence of length T the complexity is $O(T \cdot (H \cdot D \cdot w + H \cdot C))$.

(c) in zip

(d)

i. Entity level P/R/F1: 0.82/0.84/**0.83**

Token-level scores:

label	acc	prec	rec	f1
PER	0.99	0.92	0.94	0.93
ORG	0.99	0.87	0.79	0.83
LOC	0.99	0.89	0.89	0.89

MISC	0.99	0.88	0.81	0.84
O 0.99	0.99	1.00	0.99	
micro	0.99	0.98	0.98	0.98
macro	0.99	0.91	0.88	0.90
not-O	0.99	0.89	0.87	0.88

The confusion matrix tells us a lot, one of the most clear things is that even our precision is pretty good our recall is not so much generally, particularly it had poor scores for ORG and MISC especially and maybe the two got mixed up sometimes.

2 modeling limitation:

a. Long entity name which are longer than the window size are missed –

x : The **Gazzetta dello Sport** said the deal would cost Atalanta around $ 600,000 .
y*: O **ORG** ORG ORG O O O O O ORG O O O O
y': O **MISC** O ORG O O O O O O O O O O O

b. The window has finite size, while the sentence length and reference location and vary and be far ways from the entity.

x : **Woods** was among a group of 13 players at four under , including 1993 champion Billy Mayfair , who tied for second at last week 's World Series of Golf , and former U.S. Open champ Payne Stewart .
y*: **PER** O O O O O O O O O O O O O PER PER O O O O O O O O O MISC MISC MISC MISC O O O MISC MISC O PER PER O
y': **O** O O O O O O O O O O O O O PER PER O O O O O O O O O MISC MISC MISC MISC O O O MISC MISC O PER PER O

another limitation is that decisions in neighboring parts of the input are made independently from each other.

Question 2

(a)

i. W_h is size $H \times H$ and W_e is $D \times H$ plus bias. So we have $D \cdot H + H \cdot H + H$ parameters in the hidden layer of the RNN cell and additional $H \cdot C + C$ parameters for softmax.
The window-based model number of parameters is only $(2w + 1) \cdot D \cdot H + H \cdot C + H + C$. (in our case embedding size is 50, hidden size is 200 and output size is 5)

ii. At each Iteration the calculation of $e^{(t)}$ takes the same $O(w \cdot D)$ operations and $h^{(t)}$ takes $O(H \cdot D \cdot w + H \cdot H)$. For the calculation of $\hat{y}^{(t)}$ $O(H \cdot C + C)$. CE calculation is less since it's $O(C)$.
So over all since we do one forward pass across all words of a sentence of length T the complexity is $O(T \cdot (H \cdot D \cdot w + H \cdot H + H \cdot C))$.

(b)

i.

Generally speaking, that could happen for an entity with more than one word. If we improve on a single word prediction to the correct entity label this would decrease CE for that word but it could create 2 entities F_1 score wise and count more mistakes. Example:

"Barak Obama walked" (PER PER 0)

If original label was (0 0 0), then changing it (PER 0 0) would lower CE (since we are correct on the word 'Barak' now), but our F_1 score would decreasing because there is now another entity to consider where we made a mistake (missing labeling 'Obama').

ii.

Because there 2 tasks involving F_1 score – recognizing entity and labeling them – and one can contradict the other as we saw in the previous answer. This could be prevented if scoring for would consider all the data forward for each label and that is computably hard.

(d)

i.

If we did not use masking then the loss for $t \in (M - T, M)$ would be accumulated (even though it's suppose to be zero from our point of view since it doesn't represent part of the original data) and we would optimize trying to lower the CE for the null token that is not relevant.

(g)

i + ii

First problem is that each word is labeled with information about the previous parts of the sentence, but does not look forward (unlike the window-based model). This could be solved using bi-directional RNN.

x : Rotor Volgograd must play their next home game behind closed doors after fans hurled bottles and stones at Dynamo **Moscow** players during a 1-0 home defeat on Saturday that ended Rotor 's brief spell as league leaders .

y*: ORG ORG O O O O O O O O O O O O O O O O ORG **ORG** O O O O O O O O O O ORG O O O O O O O O

y': ORG ORG O O O O O O O O O O O O O O O O ORG **LOC** O O O O O O O O O O ORG O O O O O O O O

here if it has information from the next word 'players' the model would have label it correctly probably.

A second problem is that each label is labeled separately so an entity with many words that can be confused between PER and ORG for example each word of the phrase could decided differently. A solution that could be enforcing consist some how on labeled adjacent words, or enter as input the previous word decided label in addition to the hidden layer input.

x : May 25 Third one-day international (at Lord **'s** , London)

y*: O O O O O O O LOC **LOC** O LOC O

y': O O O O O O O LOC **O** O LOC O

here it's obvious that 's is part of the word of the location, but the decision is separated about the words.

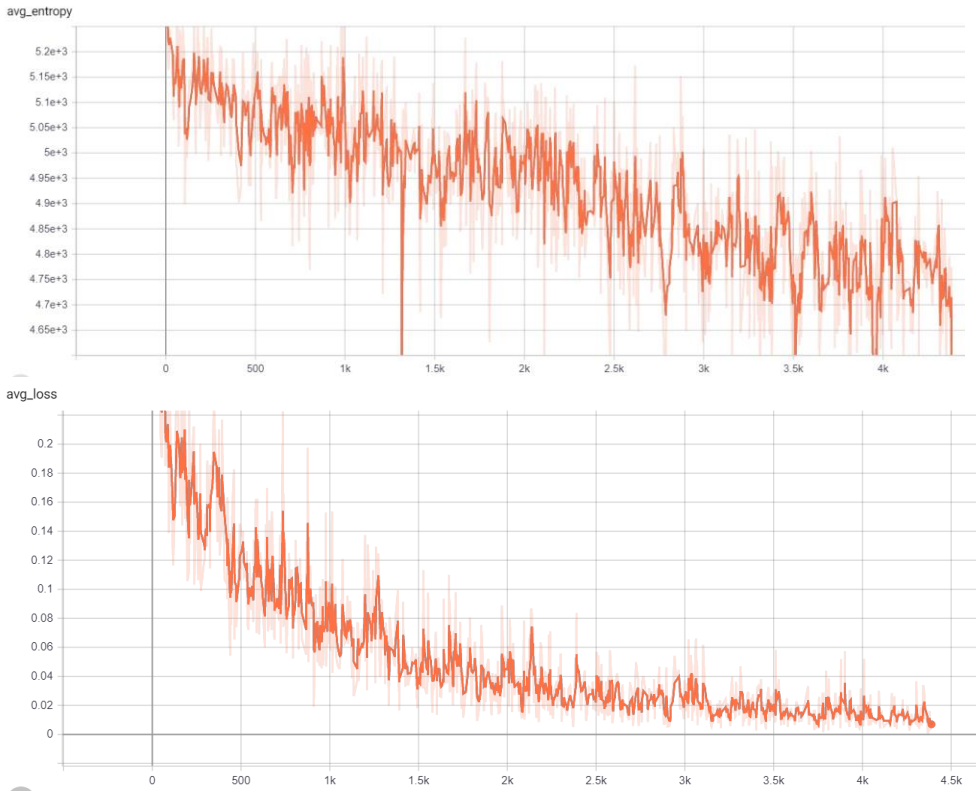
Question 3

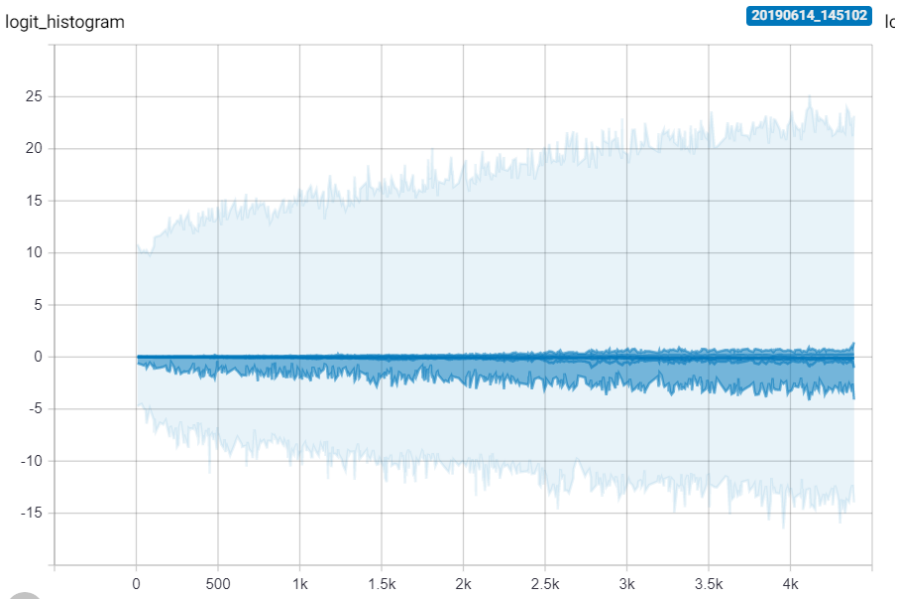
(c)

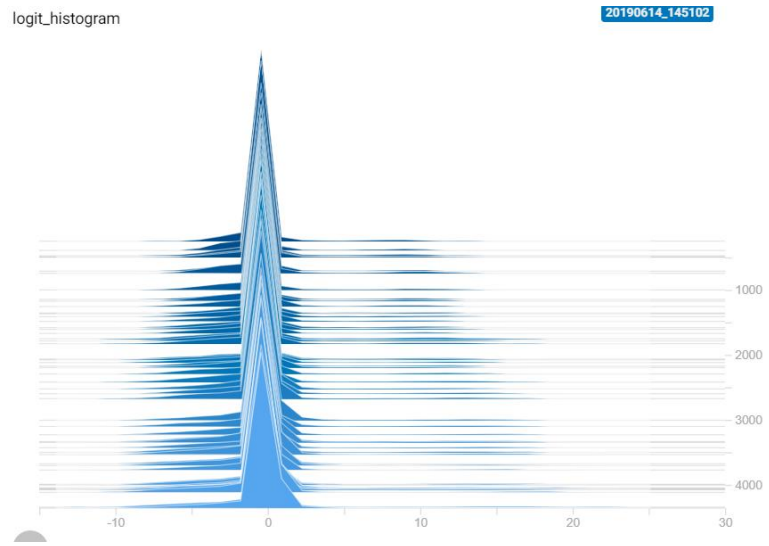
i. $-(1 * ln0) = inf$

infinity in case the value for the correct value is 0.

ii.







The entropy lowers with time meaning there is less variation in the predictions of the model. The logits values are increasing, it can be said that the values of the important features values that reach the correct class are higher with time which goes along with the decreasing of the entropy.

(d)

The model predict well the existence of an entity, with a high score for 0 and for the label micro. The model is worse for ORG and MISC labels.

For ORG labels for example it's possible to see that the probiblty classification is not so high even when the model is correct, we can see this in:

x : **Essex** , however , look certain to regain their top spot after Nasser Hussain and Peter Such gave them a firm grip on their match against Yorkshire at Headingley .

y*: **ORG** O O O O O O O O O O O PER PER O PER PER O O O O O O O O O ORG O LOC O

y': **ORG** O O O O O O O O O O O PER PER O PER PER O O O O O O O O O ORG O LOC O

p **: 0.79** 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.99 1.00

And when the model is wrong the probabilities are even lower, which makes sense given the nature of the mistake:

x : London (The Oval) : Warwickshire 195 , Surrey 429-7 (C. Lewis 80 not out , M. Butcher 70 , G. Kersey 63 , J. Ratcliffe 63 , D. Bicknell 55) .

y*: LOC O LOC LOC O O ORG O O **ORG** O O PER PER O O O O PER PER O O PER PER O O PER PER O O PER **PER** O O O

y': LOC O LOC LOC O O ORG O O PER O O PER PER O O O O PER PER O O PER PER O O PER PER O O PER PER O O O

p : 0.82 1.00 0.98 1.00 0.98 1.00 0.99 1.00 1.00 **0.59** 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00

x : **Bristol** : Gloucestershire 183 and 185-6 (J. Russell 56 not out) , Northamptonshire 190 (K. Curran 52 ; A. Smith 5-68) .

y*: **LOC** O ORG O O O O PER PER O O O O O ORG O O PER PER O O PER PER O O O

y': **ORG** O ORG O O O O PER PER O O O O O ORG O O PER PER O O PER PER O O O

p : **0.35** 0.99 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00