NLP - HW 4

Guy Tevet (305257206), Gavriel Habib (304946445)

Question 1 – A window into NER

(a) i. * Israel is 70 years old. (Israel can be a person or a country)
* Revenues of 14.5$ billion posted by Dell. (Dell can be either an organization or the CEO of some company)

ii. It is important to use features apart from the word itself to predict named entity labels, because it could help us avoid named entity ambiguity. For example: Atlantic could be either an organization or location, but if it appears after the word Pacific, then it must be a location.

iii. First feature – the word begins with a capital letter (but not in the beginning of a sentence).

Second feature – the window contains a word that relates to locations (like: near, located, etc.).

(b)
i. e^t is a vector of dimension $1\ X\ (2w+1)D$;
W is a matrix of dimension $(2w+1)D\ X\ H$;
U is a matrix of dimension $H\ X\ C$.

ii. The computational complexity of predicting one label is:
- Extracting $(2w+1)$ rows from E: $O(2w+1)$
- Matrix multiplication $e^t\ W$ and ReLU: $O((2w+1)DH)$
- Matrix multiplication $h^t\ U$ and softmax: $O(HC)$
- We don't calculate the CE complexity, because we have been asked to compute the complexity of predicting $\hat{y}_t$.

Total: $O(((2w+1)D+C)H)$

Therefore, the complexity of predicting T labels in a sentence is:
$O(((2w+1)D+C)HT)$

(d) i. Best results:

DEBUG:Token-level confusion matrix:

go\gu	PER	ORG	LOC	MISC	O
PER	2936.00	59.00	61.00	13.00	80.00
ORG	132.00	1680.00	106.00	48.00	126.00
LOC	42.00	142.00	1846.00	20.00	44.00
MISC	41.00	78.00	44.00	999.00	106.00
O	39.00	53.00	16.00	27.00	42624.00

DEBUG:Token-level scores:

label	acc	prec	rec	f1
PER	0.99	0.92	0.93	0.93
ORG	0.99	0.83	0.80	0.82
LOC	0.99	0.89	0.88	0.89
MISC	0.99	0.90	0.79	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	0.89
not-O	0.99	0.89	0.87	0.88

INFO:Entity level P/R/F1: 0.81/0.84/0.83

We can learn from the confusion matrix that the most probable errors that our model is making are:

- We predict Person for words that are Organizations.
- We predict Organization for words that are locations, and the opposite.

That is reasonable because there are examples of entity ambiguity for those pairs. For example: the word 'Atlantic' could be a location and an organization, and it depends on the full context that may not be sure enough from a small window around the word.

ii. **First limitation of the model -** small window size, that prevent the model understanding the full context of the sentence.

Examples (Dinamo is an organization and not a person):

x : Red Star (Yugoslavia) beat Dinamo (Russia) 92-90 (halftime
y*: ORG ORG O LOC O O ORG O LOC O O O O
y': ORG ORG O LOC O O PER O LOC O O O O

x : SOCCER - ROMANIA BEAT LITHUANIA IN UNDER-21 MATCH .
y*: O O LOC O LOC O O O O
y': O O LOC O ORG O O O O

x : Results from the U.S. Open Tennis Championships at the National Tennis Centre on Friday (prefix number denotes seeding) :
y*: O O O MISC MISC MISC MISC O O LOC LOC LOC O O O O O O O O O
y': O O O MISC MISC MISC MISC O O ORG LOC LOC O O O O O O O O O

Second limitation of the model - limited context window. We would like to use a model that has unlimited context window (therefore – has memory; like RNN). The reason is that sometimes, a prediction of word in some place in the sentence is defined by some (very) far word in the sentence.

Example (The word Washington is a person and not a location):

x : O'Brien , a winner two weeks ago in New Haven for his first pro title , served for the match at 5-4 in the third set before Washington came charging back .
y*: PER O O O O O O O LOC LOC O O O O O O O O O O O O O O O O O O O O PER O O O O
y': PER O O O O O O O LOC LOC O O O O O O O O O O O O O O O O O O O O LOC O O O O

Question 2 – RNNs for NER

(a)

i. RNN has H^2 more parameters than the window-based model (because of the matrix W_h).

ii. The computational complexity of predicting one label is:

- Extracting one row from E: O(1)
- Calculating h^t: O$(H(D+H))$
- Matrix multiplication $h^t\ U$ and softmax: O(HC)

Total: O$(H(D+H+C))$

Therefore, the complexity of predicting T labels in a sentence is:
O$(HT(D+H+C))$

(b)

i. Considering a single word in the train set with the ground truth tag **LOC** , and the following change of the output probability vector:

$$\begin{bmatrix} \Pr(O) = 0.25 \\ \Pr(LOC) = 0.3 \\ \Pr(PER) = 0.25 \\ \Pr(ORG) = 0.2 \end{bmatrix} \rightarrow \begin{bmatrix} \Pr(O) = 0.5 \\ \Pr(LOC) = 0.4 \\ \Pr(PER) = 0.05 \\ \Pr(ORG) = 0.05 \end{bmatrix}$$

- The final prediction changed from **LOC** to **O** which decreases the precision and hence decreases F1.
- although the prediction was changed, the probability of the correct tag itself was raise, which means that the cross-entropy loss also decreases.

ii.It is difficult to directly optimize F1 because this function is not continuous and therefore not differentiable.

(d)

i. If we won't use masking, the loss and gradient updates would change. This can cause an increase in the loss, because the network may predict some label on the places outside the mask, and it would affect the gradients as well. Of course, this is an unnecessary effect because the change is artificial. If we would use masking, it would solve the problem because it zeros the loss on those NULL labels.

(f) Results of the model after training:

DEBUG:Token-level confusion matrix:

go\gu	PER	ORG	LOC	MISC	O
PER	2971.00	41.00	74.00	11.00	52.00
ORG	137.00	1667.00	101.00	53.00	134.00
LOC	51.00	101.00	1885.00	20.00	37.00
MISC	42.00	51.00	42.00	1023.00	110.00
O	51.00	55.00	20.00	30.00	42603.00

DEBUG:Token-level scores:

label	acc	prec	rec	f1
PER	0.99	0.91	0.94	0.93
ORG	0.99	0.87	0.80	0.83
LOC	0.99	0.89	0.90	0.89
MISC	0.99	0.90	0.81	0.85
O	0.99	0.99	1.00	0.99
micro	0.99	0.98	0.98	0.98
macro	0.99	0.91	0.89	0.90
not-O	0.99	0.90	0.88	0.89

INFO:Entity level P/R/F1: 0.83/0.86/0.84

(g) **First limitation of the model –** RNN knows only the past but not the future. It can prevent the model understanding the full context of the sentence. Suggested solution: bidirectional RNN.

Example (Brown Deer is not a person, but the name of the park):

```
x : 6,739-yard Brown Deer Park Golf Course after the first round
y*: O       LOC  LOC  LOC  LOC  LOC   O    O  O    O
y': O       PER  PER  LOC  MISC O     O    O  O    O
```

Second limitation of the model – The model does not enforce adjacent tokens to have the same tag. We would suggest that the model would understand which words are belong to the same phrase, and therefore it will have the ability to enforce the same tag for all of them.

Example ("berlin grand prix" is a phrase):

```
x : ATHLETICS - BERLIN GRAND PRIX RESULTS .
y*: O       O MISC  MISC  MISC O      O
y': O       O LOC   MISC  MISC O      O
```

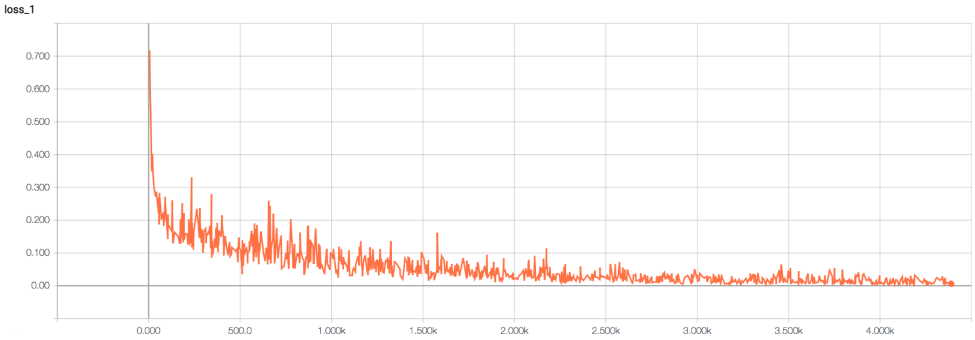
Question 3 – GRUs and TensorBoard

(c)

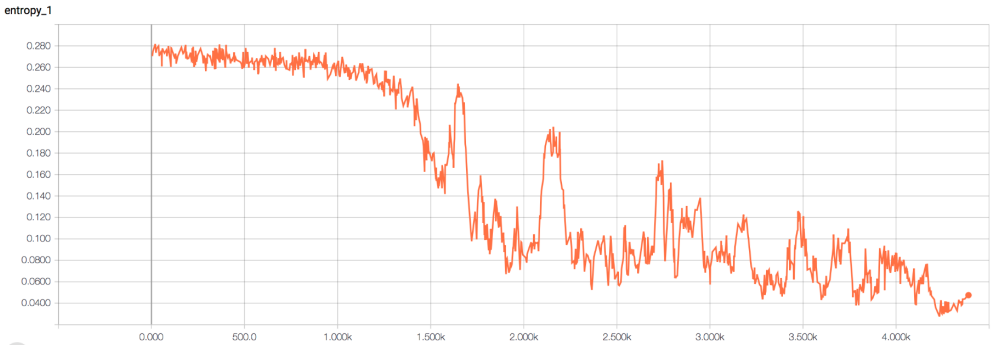
i. The maximum entropy value possible for a single timestep prediction is $\frac{1}{e}$. This is by calculation the derivative and comparing it to 0.

ii. Analyzing the training graphs:

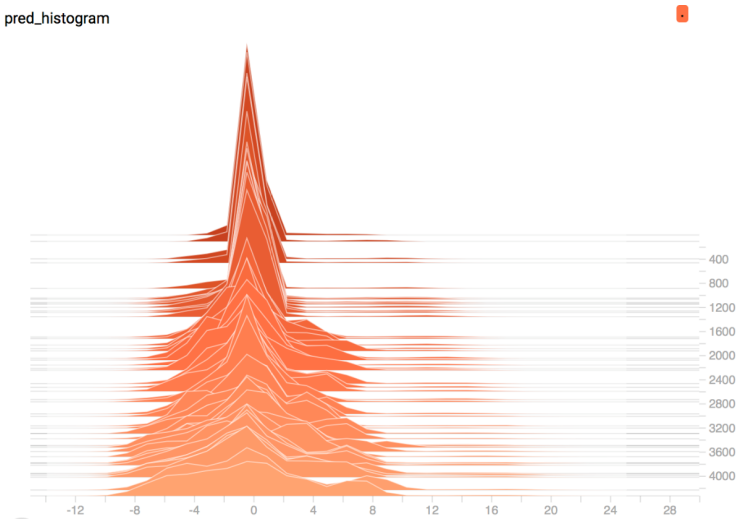
- Loss: In general, this is a monotonic decreasing graph that converges to some local minima value (close to 0). Though, the graph is very noisy, and that is because the loss is calculated on a small batch each iteration, and not on the whole dataset.



- Average entropy: This is also a decreasing graph in general, but it's not a monotonic graph (there are some places that it is even increasing). It can be explained by the fact that the loss and the entropy are correlated but not equal. Therefore, some iterations that minimize the loss not necessarily minimize the entropy. We can also see that the graph is noisy (probably because of the same batch reason).



- Predictions histogram: As we move forward in the iterations, we can see that the prediction values are spreading in a larger range and distributing more uniformly on that range.



(d) Here are the results of F1 scores and confusion matrices:

go\gu	PER	ORG	LOC	MISC	O
PER	2889.00	102.00	46.00	25.00	87.00
ORG	80.00	1713.00	63.00	138.00	98.00
LOC	28.00	101.00	1876.00	54.00	35.00
MISC	25.00	41.00	31.00	1083.00	88.00
O	20.00	64.00	11.00	41.00	42623.00

2018-05-12 17:56:02,749:DEBUG: Token-level scores:

label	acc	prec	rec	f1
PER	0.99	0.95	0.92	0.93
ORG	0.99	0.85	0.82	0.83
LOC	0.99	0.93	0.90	0.91
MISC	0.99	0.81	0.85	0.83
O	0.99	0.99	1.00	0.99
micro	0.99	0.98	0.98	0.98
macro	0.99	0.90	0.90	0.90
not-O	0.99	0.90	0.88	0.89

2018-05-12 17:56:02,749:INFO: Entity level P/R/F1: 0.85/0.86/0.85

We can see that the labels that the model is struggling to predict well are: Location and Organization (it sometimes swaps between those two). In addition: Person and Organization; MISC and Organization.

An example in which our model had some predictions errors:

```
x : Two  die  as   New  Hampshire motel explodes and  burns .
y*: O    O    O    LOC  LOC      O     O       O    O    O
y': O    O    O    O    ORG      O     O       O    O    O
p : 1.00 1.00 1.00 0.41 0.97     1.00  1.00    1.00 1.00 1.00
```

Another funny example is the word “New” that the model is struggling to predict without looking at the future. This is because this word is usually not an entity, but in some cases it appears to be the beginning of a location phrase (such as “New Hampshire” above).

Nevertheless, the model predicts well other labels such as: O (not entity) and Person, as we can see from the F1 scores above.