Ex4 - NLP

Yoni hornstein

Noa Koren

Udi rot

Question 1

a. q1
   i. Two examples:
      1. Philip Morris decided to stop manufacturing cigarettes - it isn't clear whether Philip Morris is a person or a company
      2. Jane Bordeaux released a new song - jane bordeaux is a band but may be considered a person.
   ii. Named entities are more likely to be rare words, so more features are needed to predict the label correctly
   iii. The features:
      1. Casing - named entities are supposed to be cased
      2. POS features

b. Dimensions:
   i. $e^{(t)} - 1 \times (2w+1)d$
   ii. $W - (2w+1)d \times H$
   iii. $U - H \times C$

   Complexity:

   $e^{(t)}$is consisted of (2w+1)*d simple elements, so it takes O((2w+1)d) to compute.

   Computation of h takes O((2w+1)*d*h + h) because of the addition of b1, and for the same reason the computation of y takes O(H*C + C).

c. Code
d. Code
e. Analysis
   i. The model got 82% F1. The table:

go\gu	PER	ORG	LOC	MISC	O
PER	2936	62	59	12	80
ORG	127	1681	97	49	138

LOC	41	161	1814	19	59
MISC	42	75	46	995	110
O	35	55	16	30	42623

By looking at the confusion table, we can tell that ORG is the hardest label to predict - it is too often predicted as other labels, mostly PER.

Ii. blem tof the window model is that it cannot use previous or future predictions to predict a label based on the labels of the word before it. For example, for the sentence "Ruehe planned to meet his Israeli counterpart Yitzhak Mordechai and Israeli President Ezer Weizman , the ministry said.", Ruehe is predicted to be an organization, although the presence of 'his' in the sentence indicates it is a person.

Also, it would be hard for the model to identify entities whose name is longer than the window - for example, not all tokens in "Test and County Cricket Board" were recognized as ORG - only "County Cricket Board".

Question 2

2. a. i.

The parameters in the RNN model are W_e and W_h, while in window model it is W.

W is $(2w+1)D \times H$ dimensional (which is also the number of parameters), while W_e is $D \times H$ dimensional, and in addition W_h is $H \times H$ -dimensional.

ii.

for labels predicting for a sentence of length T for the RNN model, we need to compute $e^{(t)}$, $h^{(t)}$ and $\hat{y}^{(t)}$.

For $e^{(t)}$ we need to perform $O(D)$ operations (as we multiply **one hot vector** with size V, with a $V \times D$ dimensional matrix).

For $h^{(t)}$ we need to perform $O(H^2 + DH + H)$ operations (for two multiplications, and an addition, and sigmoid function on H-dimensional vector).

And for $\hat{y}^{(t)}$ we need to perform $O(HC + C)$ operations (for one multiplication, and an addition, and softmax on C-dimensional vector).

So, the total complexity is $O\big(T * (D + H^2 + DH + H + HC + C)\big) = O(T * \big(H(D + H) + H + D + H(C + 1)\big) = O\left(T * \big((H + 1) * (D + H) + \mathrm{H} * (\mathrm{C} + 1)\big)\right) \sim \boldsymbol{O(T * H(D + H))}$

As we know that C = 5 (and generally is a small number).

2.b. i.

Let's look at the example:

Avi Cohen/MISC is a professor at TAU/MISC.

If we predicted Avi Cohen/O is a teacher at TAU/O and then in the next step:

Avi Cohen/Misc is a teacher at TAU/O

The loss entropy is decreased as we predicted 'Avi Cohen' correctly,

But the precision got lower as we add another wrong entity prediction, and so is F1 score.

ii.

it's hard to directly optimize F1 because F1 score is non convex and not differentiable.

d. i. Masking is used to zero the influence that comes from the padding 0-vectors, as they are used only for padding (and are not supposed to add any new information). Multiplication the loss elements with the masking vector actually zeros the irrelevant elements and keeps only the relevant data.

Without using the masking, the loss will include the prediction of the zero labels, which of course are meaningless as they are not really exist. The paddings vector gradients will affect the resulted parameters.

f. – Trained our model – (F1 = 0.85 %).

g.

the First limitation is that words from the same phrase do not have to have the same Tag.

Example:

x : **CITIC Pacific** is a major Hong Kong-listed company focusing on infrastruture , trading , distribution and property , with 28 percent of its 1995 profits coming from telecoms .

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

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

The second limitation is that RNN sees only the past - but not the future. So it loses some of the sentence context.

Example:

x : **Longyear is a town** in mourning , a close-knit community that has been shattered .

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

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

Here, 'is a town' appears after **Longyear** and it makes sense that it may clarify that Longyear is a location ('LOC').

In order to solve this, we can use birnn, which is bi-directional RNN.

Question 3

c.i.

It is known that the entropy is the highest for an uniform distribution, a.k.a all the probabilities for tags are equal.

In this case, entropy = $-\sum \frac{1}{c} \ln\left(\frac{1}{c}\right) = -\ln\left(\frac{1}{c}\right) = \ln(c)$. (c is the number of classes).

ii.

The average entropy graph is a decreasing graph. First, the initial parameters (weights) are the parameters that fits to uniform distribution, which leads to the highest entropy, so it makes sense. In addition, lower entropy indicates that the model predictions are more specific – so if we trained our network more, it gets more certain. As we said, the entropy graph is indeed a decreasing graph, which fits the theory.

The loss graph is a decreasing graph as well, falling to zero. This fact together with the behavior of the average entropy graph tells us that the predictions indeed gets more specific – but also true.

In the logits histogram we can see that for a bigger n, the logits are spreading in a larger range. This also indicates that the entropy is indeed lower.

d.

The model got 85% F1, the confusion table:

go\gu	PER	ORG	LOC	MISC	O
PER	2959	50	43	12	85

ORG	120	1714	101	64	93
LOC	28	103	1918	16	29
MISC	35	50	65	1005	113
O	33	87	21	22	42596

The confusion table implies the model handles PER and LOC well but has difficulties handling ORG and MISC. It often predicts MISC as O and ORG as PER. For example, Titanic is predicted to be PER, ORG and O, but never MISC.

Like previous models, in the sentence “Ruehe planned to meet his Israeli counterpart Yitzhak Mordechai and Israeli President Ezer Weizman , the ministry said.” Ruehe is predicted to be ORG and not PER, because it doesn’t consider ‘his’ while predicting it. The model doesn’t consider the future, so the sentence context isn’t fully understandable sometimes.