ADVANCED METHODS IN NLP
EXERCISE #4 SOLUTION

Uri Avron [uriavron@gmail.com] [308046994]
Ofri Kleinfeld [ofrik@mail.tau.ac.il] [302893680]
Ido Calman [calman.ido@gmail.com] [308353499]

May 21, 2018

Question 1

(a.i)

(1) Georgia is so beautiful. (Georgia may be either the state (location) or a person)

(2) Michael Jordan is considered the best basketball player of all time. (Jordan maybe the country, the previous word is the key here)

(a.ii) It is important to use features apart from the word itself because the context is relevant, as we see in the two examples above.

(a.iii)

(1) The words around the current word.

(2) The structure of the word (i.e. ends with a dot, all upper-case)

(b.i) $\mathbf{e}^{(t)}$is $1 \times (2w+1)D$, W is $(2w+1)D \times H$, U is $H \times C$.

(b.ii) Computing $\mathbf{e}^{(t)}$costs $O\left((2w+1)D\right)$, computing $\mathbf{h}^{(t)}$costs $O\left((2w+1)DH\right)$, computing $\hat{\mathbf{y}}^{(t)}$ costs $O(HC)$. Overall, it costs $O\left(T\left((2w+1)D + (2w+1)DH + HC\right)\right)$ to predict lables for a sentence of size T.

(d.i) BestF_1 score is 0.82. The token-level confusion matrix is

go\gu	PER	ORG	LOC	MISC	O
PER	2911	67	82	14	75
ORG	117	1685	98	63	129
LOC	32	178	1817	26	41
MISC	28	87	31	1010	112
O	30	51	17	30	42631

We can infer from the confusion matrix that the model has difficulties to distinguish between LOC and ORG.
Moreover, the model has difficulties to identify ORG in general - we can see that from the matching row in the confusion matrix.
The rows in the confusion matrix are the true labels while the columns are the predicted ones - meaning large number of mistake in ORG row will result in low recall for this label.

(d.ii)

(1) We used a window of size 1, which does not capture long-term contexts between words. As we described in (b.ii), the computational complexity depends on the window size so a bigger window may increase heavily the training and testing time.

(2) Our model depends only on words inside the window, and does not take into account other potentially indicative information in the sentence to encode as features. Indicative information may be the structure of a word (upper-case letters, suffixes etc.), previous tags and more. (Things we talked about in the log linear models for tagging classes)

(examples)

1. In “My girls’ names are Dana, Dina and Donna”, the model predicted successfully that Dana and Dina are persons, but Donna was identified as organization. This supports the limitation of the window of size 1 and that it does not look on previous tags. We believe that if while predicting the tag for “Donna” the model could have used the information about the previous two PER tags - it might have been predicting also “Donna” as a PER.

2. Another example could be “My best friend’s name is Jordan”.
The model tagges “Jordan” as LOC beacuse it only has information about the previous word “is”.
We believe that increasing the window size even to 2-4 words could probably make the model overcome this adversary example.

Question 2

(a.i) Number of parameters for the RNN model: $VD+H^2+DH+H+HC+C$
Number of parameters for the window-based model: $VD+(2w+1)DH+H+HC+C$
Thus, the difference between the models is the difference between H^2+DH and $(2w+1)DH$. i.e it dependes on the window size w (RNN will have more parameters when $\frac{H}{2D} > w$)

(a.ii) Computing $\mathbf{e}^{(t)}$costs $O\left(D\right)$, computing $\mathbf{h}^{(t)}$costs $O\left(DH+H^2\right)$, computing $\hat{\mathbf{y}}^{(t)}$ costs $O(HC+C)$. Overall, it costs $O\left(TH\left(D+H+C\right)\right)$ to predict lables for a sentence of length T.

(b.i) We can consider an example scenario:
Trying to tag the following sentence: “The/O kingdom/LOC palace/LOC” and moving from initial tags “The/O kingdom/O palace/O” to the following decision - “The/O kingdom/O palace/LOC”.
We have decreased our cross entropy loss because we are tagging another token correctly, but our entity level F1 score would also decrease because now we have predicted another entity incorrectly (precision decreases and recall remains the same)

(b.ii) It is difficult to directly optimize F1 score because it is non-convex, and worse than that, not even differentiable. That is why we are using the cross-entropy loss as a surrogate/approximation function that is closed to the F1 score we really want to minimize (Although it is not always the case as seen in the previous section).

(d.i) If we did not use masking, the loss can only get larger, and the gradient may point to a direction which is not in our interests, because it takes into consideration the padded dimensions. The masking solves the problem as it ignores the loss in the padded directions.

(g.i) Two modeling limitation of the RNN model:

1. One limitation of this model is that although it considers the entire sentence, it consideres only the words themselves without any other potentially indicative features of the words or the entire sentence.
   As a supporting example, we can see the following sentence from the development set: "'We had two very strong economic reports', said David Shulman , Salomon Bros".
   The word "Bros" was tagged as PER instead of ORG. It is reasonable to believe that considering the predicted tags for the previous words "David Shulman" (which were tagged correctly as PER) and the preticted tag for the word "Salomon" (which was tagged correctly as ORG), could have help the model to infer that it is indeed more reasonable to tag "Bros" as ORG rather than PER.

2. The model only uses past parts of the sentence. But, we are examining the entire sentence at once and our goal is to predict NER for the entire sentence (sequence to sequence). So, there isn't really a limitation to use only past parts of the sentence and not to icnlude also later parts of the sentence, like we actually did with the window tagger.
   As a supporting example, we can see the following sentence from the development set: "I will in fact move to Kecskemet (site of Petofi printing house)".
   The word "Petofi" was tagged as LOC instead of ORG. It is reasonable to believe that using the two words that comes after "Petofi" - "printing house", the model could have concluded that "Petofi" is indeed an ORG rather than LOC.

(g.ii) Suggested extentions for the model:

1. We can use feature rich models that takes into account also things like previous tagging decisions, the length of the current word, does the current word consists only of captial letters, etc. Similar to what we did with log linear models for POS tagging in class.

2. In order to use later parts of the sentence, we can also roll them through the network similar to the way we are rolling past parts of the sentence. One example of an implementation of this kind of approch is Bi-LSTM networks.

Question 3

(c.i) The maximun entropy value possible for a single timestamp prediction is $\ln(C)$, where C is the number of classes available for prediction.
This follows the case where all the possible tages are equally probable. Thus the entropy will be: $-\sum_{i=1}^{i=C} \frac{1}{C} \ln(\frac{1}{C}) = -C\frac{1}{C} \ln(\frac{1}{C}) = \ln(C)$

(c.ii) The average entorpy over time graph tells us how much our prediction vector values are concentrated - as explained in previous section, when all the vector entries are equally probable we will get high entorpy value, while when the distribution mass is concentrated in only one entry we will get entropy value of 0 which is the lowest possible value.
The decrease in entropy over time ensures us that our prediction vectors gets concentrated towards one specific or few specific entries. Together with the decrease of the loss fucntion over time we can conclude that the prediction vectors are getting concentrated towards the **right vector entries** (i.e the true class of the vector).

(d) The model predicts well the labels PER and LOC but struggles more with ORG and MISC labels - as seen by the F1 scores for the last two labels. (F1 score of 85% compared to 93% and 92% for PER and LOC). The other possible tags are also being predicted well with more than 90% F1 scores.
We can see that the model is still struggling to differentiate between some of the labels where the rest of the sentence is highly indicative to the current tagging decision.
For example in: “The Australian had upset Jansher ’s rhythm with his mixture of gamesmanship and fluent stroke-making but eventually succumbed 15-7 17-15 14-15 15-8”
The model predicted “Jansher” as LOC instead of PER. We believe that if the model could take into account the rest of the sentece, even just the “’s” word (parsed as a different word because there is a space between them), it would have tag “Jansher” as PER instead of LOC.