

# Assignment 3 writeup

Thomas Lu

## 1 Problem 1

(a)

- (i)
  - Belgium beat Brazil 2-1 in today's World Cup match. (Does "Belgium" refer to the country, a location, or the Belgian national team, an organization? Similarly for Brazil.)
  - If you want the best diamonds, go to Tiffany. (Is "Tiffany" a person, or does it refer to the jewelry store, which is a location?)
- (ii) The word itself might appear in different sentences as different types of entities, and the only way to distinguish which type of entity it is representing is through examining the context.
- (iii)
  - If a noun does not have a dependent determinant (the/a/an/etc.), it is more likely to be a named entity.
  - A noun followed by an appositive phrase (a descriptive phrase contained between two commas) is more likely to be a named entity, especially if the appositive phrase doesn't contain any specifying pronouns (who, which, etc.).

(b)

- (i)  $e^{(t)}$  will be  $1 \times D(2w + 1)$ ,  $W$  will be  $D(2w + 1) \times H$ , and  $U$  will be  $H \times C = H \times 5$ .
- (ii) The matrix multiplications (which dominate the asymptotic computational complexity) required are:
  - $T$  multiplications  $e^{(t)}W$  of a  $1 \times D(2w + 1)$  vector by a  $D(2w + 1) \times H$  matrix, which contributes  $O(TD(2w + 1)H) = O(TDwH)$  complexity.
  - $T$  multiplications  $h^{(t)}U$  of a  $1 \times H$  vector by a  $H \times C$  matrix, which contributes  $O(THC)$  complexity. Taking  $C = 5$  to be constant, this is  $O(TH)$ .

Thus the total complexity is  $O(TH(Dw + 1)) = O(THDw)$ .

(d)

- (i) The best dev entity-level F1 scores achieved was 0.84. The token-level confusion matrix was:

go\gu	PER	ORG	LOC	MISC	0
PER	2947.00	33.00	97.00	10.00	62.00
ORG	129.00	1659.00	122.00	54.00	128.00
LOC	34.00	89.00	1895.00	14.00	62.00
MISC	41.00	59.00	55.00	1006.00	107.00
0	39.00	44.00	20.00	37.00	42619.00

From the confusion matrix, we can see that the model is commonly mispredicting instances of ORG as PER, LOC, or O, and commonly mispredicts instances of MISC as O. (This is corroborated by the token-level scores, which show low recall for the ORG and MISC categories.)

- (ii)
  - Because of the short window, the model has trouble detecting longer entity names, for example, “Duke of Norfolk’s XI” in the below:

```
x : May 15 v Duke of Norfolk 's XI ( at Arundel )
y*: 0  0  0 ORG ORG ORG      ORG ORG 0 0 LOC  0
y': 0  0  0 ORG 0  LOC      0  ORG 0 0 LOC  0
```

This was particularly harmful to organizations as organizations often have longer names than other named entities.

- If some prefix of an entity name can be the name of a different type of entity (e.g. “Hong Kong” is a LOC, but “Hong Kong Open” is a MISC), and only this prefix appears in the window, the model can make a mistake, as with the example below:

```
x : SQUASH - HONG KONG OPEN QUARTER-FINAL RESULTS .
y*: 0      0 MISC MISC MISC 0      0      0
y': 0      0 LOC  MISC MISC 0      0      0
```

## 2 Problem 2

(a)

- (i)  $b_1, b_2, U$  do not change,  $W_e$  has  $DH$  parameters instead of  $(2w + 1)DH$ , we have a new parameter  $W_h$  with  $H^2$  parameters. Thus the RNN has  $H^2 - 2wDH$  more parameters.
- (ii) At each time step  $t$ , the model must perform three matrix multiplications:
  - $h^{(t-1)}W_h$ , which is a  $1 \times H$  vector multiplied by a  $H \times H$  matrix, for complexity  $O(H^2)$
  - $e^{(t)}W_e$ , which is a  $1 \times D$  vector multiplied by a  $D \times H$  matrix, for complexity  $O(DH)$
  - $h^{(t)}U$ , which is a  $1 \times H$  vector multiplied by a  $H \times C$  matrix, for complexity  $O(HC)$

Thus the total complexity is  $O(T(H^2 + DH + HC)) = O(TH(H + D + C))$ .

(b)

- (i) Suppose the distribution is very skewed, with 99.9% of samples being negative and only 0.1% being positive. Suppose we were simply provided a constant hypothesis  $p$ ;  $p = 0.001$  (leading to a negative 0 for every sample) would minimize the cross-entropy, but would lead to a F1 score of 0, while  $p = 0.5$  (and resolving this to a positive or negative prediction with equal probability) would lead to a higher cross-entropy but a F1 score of 0.5.
- (ii) The F1 score is not differentiable; it shows a step change when any single prediction crosses the 0.5 threshold.

(d)

If we didn’t use masking, our output would continue producing outputs  $\hat{y}^{(t)} = \text{softmax}(\sigma(W_h h_{t-1} + W_e x_0 + b))$  for  $T < t \leq M$  for each sentence in a batch with length  $T < M$ , where  $M$  is the maximum length across sentences in the batch and  $x_0$  is the embedding of the null token. We would require our model to learn that

when the null token is passed in, it should return the null output, regardless of state. Masking allows our model to ignore this restriction.

(g)

- (i)

- The RNN model has limited ability to look into the future, since it only uses a window size of 1. Thus (similarly to the window-based model) it can get tripped up on entities who have prefixes that are a different type of entity, such as “Hong Kong Open” in the below example:

```
x : SQUASH - HONG KONG OPEN QUARTER-FINAL RESULTS .
y*: 0      0 MISC MISC MISC 0      0      0
y': 0      0 LOC  MISC MISC 0      0      0
```

- The model also has trouble with named entities that come at the beginning of sentences. Oftentimes locations and organizations are named after people, so there isn’t really any helpful information in the words themselves, and the model has no other information to work on. Example:

```
x : Apic Yamada - 6mth parent forecast .
y*: ORG  ORG   0 0   0      0      0
y': PER  PER   0 0   0      0      0
```

- (ii) Both of these issues can be addressed by allowing the model to see farther into the future. There are a number of ways this can be done. Two examples are:
  - Increase the window size.
  - Make the RNN a bidirectional RNN.