CS 224N: Homework 3

Chen Qian

Question 1

(a) i.
(1) How much do you know about Washington? (here Washington can either refer to the president or the state)
(2) You should ask Google for help. (here Google can either refer to the company or the search engine)

ii. To include the meaning of context, which could somewhat eliminate the ambiguity of a word.

iii. Constituency; Dependency; Location of the word.

(b)
i.
$e^{(t)}$: 1 * (2w + 1) D
W: (2w + 1) D * H
U: H * C.
ii.
For each window, the computational complexity is:

$$\begin{aligned} T &= T(\text{compute} \quad e^t) + T(\text{compute} \quad h^t) + T(\text{compute} \quad \hat{y}^t) \\ &= O((2w+1)D) + O((2w+1)D * H) + O(H * C) \\ &= O(wD * H + H * C). \end{aligned}$$

Notice that x^t is a one-hot vector, so $x^t L$ takes $O(D)$ instead of $O(V * D)$ time.
Thus, for a sentence of length T, the computational complexity is $O(T(wD * H + H * C))$.

(d)
i.

go\gu	PER	ORG	LOC	MISC	O
PER	2940.00	69.00	64.00	12.00	64.00
ORG	129.00	1683.00	83.00	63.00	134.00
LOC	41.00	116.00	1859.00	33.00	45.00
MISC	38.00	64.00	31.00	1024.00	111.00
O	39.00	52.00	16.00	27.00	42625.00

Figure 1: token-level confusion matrix

The best F1 score is 84%. The confusion matrix is shown in Fig. 1. According to the confusion matrix, we claim that Person entity is likely to be classified as Organization, and Organization and Location are likely to be misclassified as each other.
ii.

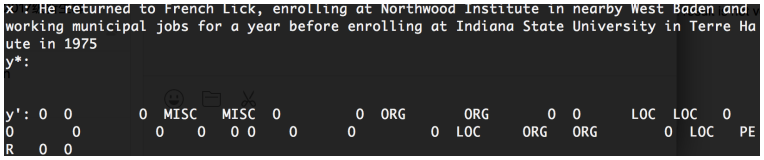

Figure 2: The example for limitation 1

Limitation 1: The performance of this window-based model is largely impacted by the size of window. For words in sentences longer than the window size, the classification result is not very accurate. The example is shown in Fig. 2.

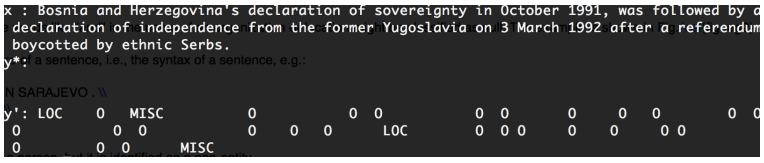

Figure 3: The example for limitation 2

Limitation 2: Words are examined individually, so some words like "and" in the name of an organization or location might be classified as null. The example is shown in Fig. 3.
Limitation 3: The model somewhat ignores the structure of a sentence, i.e., the syntax of a sentence, e.g.:
x : ATHLETICS - HARRISON , EDWARDS TO MEET IN SARAJEVO .
y*: O O PER O PER O O O LOC O
y': O O O O ORG O O O LOC O

Due to the structure/syntax, Harrison is and should be a person, but it is identified as a non-entity.

Question 2

(a)
i. In RNN, we add a new parameter matrix W_h, whose dimension is $H*H$. Thus, RNN has $H*H$ parameters than window-based model.

ii. Recalling that RNN at a time takes one word as the input, and outputs the prediction. Thus, we only need to compute the computational complexity of one turn, and the total time is just the result times T. The time for one turn is:

$$\begin{aligned} T &= T(\text{compute} \quad e^t) + T(\text{compute} \quad h^t) + T(\text{compute} \quad \hat{y}^t) \\ &= O(D) + O(D*H + H*H) + O(H*C) \\ &= O(D*H + H*C + H*H). \end{aligned}$$

Hence, the total computational complexity is $O(T(D*H + H*C + H*H))$.

(b)

i. For example, if we have two words x and y, separately of Person label and null label. Initially, the probability of classifying x as Person is 0.3, which is the maximal among all categories, and the probability of classifying y as null is also 0.3, which is also the maximal. In this case, the F1 score is $\frac{2}{1+1} = 1$. Then, by some way, we keep the probability of classifying x as Person being 0.3, but no longer the maximal among all categories, and increase the the probability of classifying y as null to 0.5. In the new case, cross-entropy loss decreases, but the F1 score drops to $\frac{2}{\frac{1}{0}+\frac{1}{0}}$.

ii. Because F1 score is not continuous. Thus, if we use F1 score as the loss function, then we cannot compute the gradients of parameters.

(d)
The loss will be larger because we take more items into account, and gradient updates will be noisy because we compute gradients on some meaningless data. If masking is applied, for $t > T$. $m^{(t)}$ is 0. Thus, when computing loss and gradients, complementary words will not be considered, so it solves the problem.

(g)
i.

```
x : Belgrade 's main airport Aerodrom Beograd will be closed to traffic for runway maintenance and modernization from September 21 to 26 , the Yugoslav news agency Tanjug reported
late on Friday .
y*: LOC   0 0  0   LOC   LOC   0  0 0   0 0   0  0   0     0  0     0  0    0 0 0 0 0  MISC  0  0   ORG  0
0  0 0   0
y': LOC   0 0  0   ORG   ORG   0  0 0   0 0   0  0   0     0  0     0  0    0 0 0 0 0  MISC  0  0   ORG  0
0  0 0   0
```

Figure 4: The example for limitation 2

```
x : The Southeast Asian gold market was more or less a photo-fit picture of the previous week 's position with activity slow and bullion prices trapped in a well-worn range
awaiting a seasonal upturn in demand and prices into the fourth quarter .
y*: 0  MISC   MISC 0  0   0  0  0 0  0 0    0   0 0 0   0  0 0   0  0   0  0 0   0   0   0 0 0    0  0
0 0    0   0 0   0  0   0  0 0   0  | 0
y': 0  LOC    LOC  0  0   0  0  0 0  0 0    0   0 0 0   0  0 0   0  0   0  0 0   0   0   0 0 0    0  0
0 0    0   0 0   0  0   0  0 0   0    0
```

Figure 5: The example for limitation 2

Limitation 1: RNN cannot take future words into consideration. The example is shown in Fig. 4.
Limitation 2: RNN suffers from gradients vanishing. The example is shown in Fig. 5.
ii.
Solution 1: Use bi-directional network instead.
Solution 2: Use GRU units instead.

1 Question 3

(a)
i.
$w_h = 1, u_h = 1, b_h = -0.5$.
ii. There are many possible solutions, let us provide one as the following:
$u_z = 1, w_z = 1, u_h = 1, w_h = 0$.

(b)
i. Assume that RNN cell could replicate the behavior, there must be the following equations:

$$u_h + b_h > 0$$
$$u_h + w_h + b_h < 0$$
$$w_h + b_h > 0$$
$$b_h < 0.$$

These 4 equations separately correspond to:
(1) h = 0, x = 1, jump to h = 1.
(2) h = 1, x = 1, jump to h = 0.
(3) h = 1, x = 0, maintain h = 1.
(4) h = 0, x = 0, maintain h = 0.
Since $b_h < 0$ and $u_h + b_h > 0$, we have $u_h > 0$. We also have $w_h + b_h > 0$, so there must be $u_h + w_h + b_h > 0$, which is contradictory to the second equation. Thus, we claim that RNN cell cannot replicate the behavior.

ii.
There are still many candidate solutions, we provide one as the following:
$b_r = 1, w_z = 1, u_z = -1, u_h = 1, w_h = -2$.

(d)

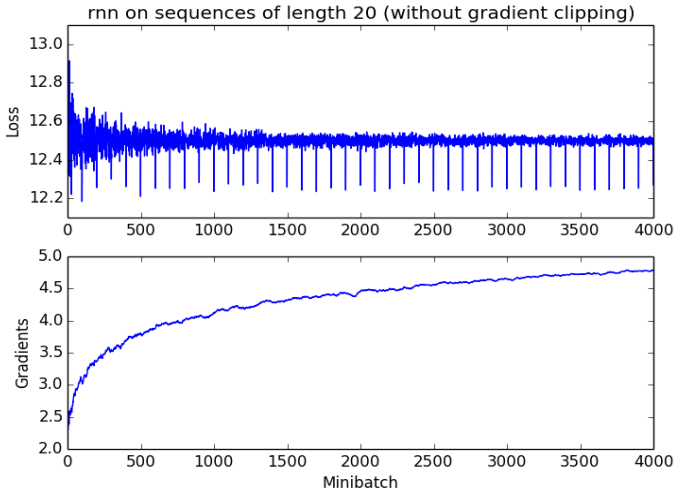

Figure 6

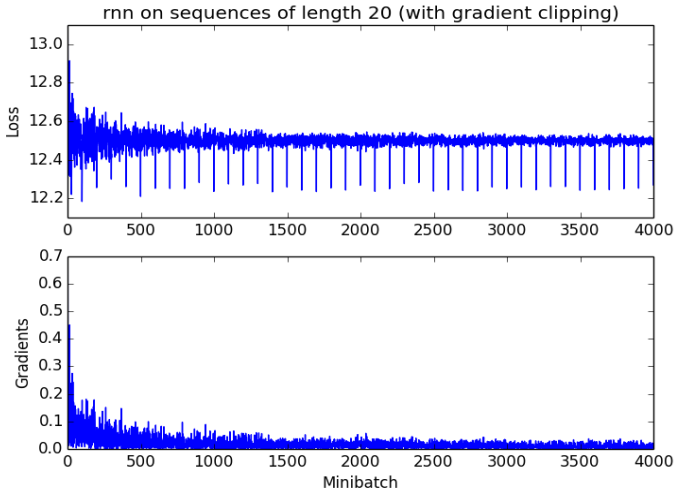

Figure 7

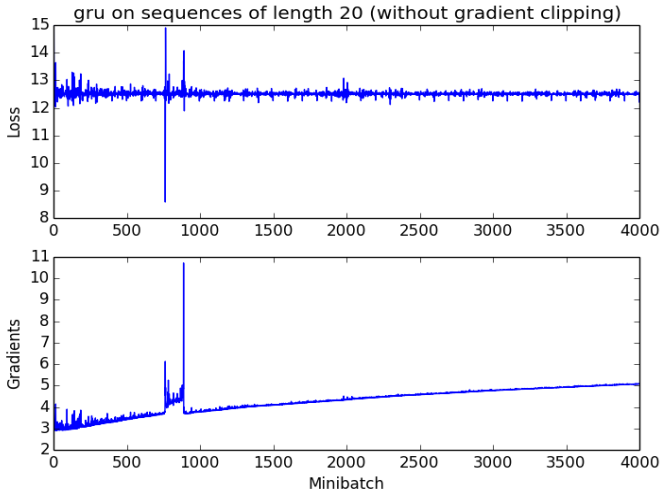

Figure 8

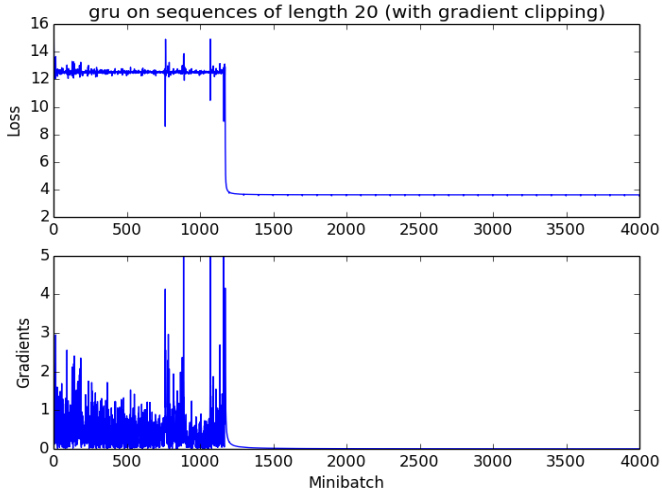

Figure 9

i. There is gradients exploding if we do not apply clipping, and clipping helps solve the problem.

ii. GRU works better than RNN. Because GRU keeps around memories to capture long distance dependencies.