CS224n = Deep Learning for NLP Stanford Course 2017

Assignment 3

1) NER Window-based

1) a)

i) example 1: The united nations, have decided to ...

$\underbrace{\text{united nations}}$ → Could be interpreted as an organisation or not an entity

example 2: Chealsea is a good $\begin{cases} \text{- man} \\ \text{- city} \end{cases}$

$\underbrace{\text{Chealsea}}$ → depending on surrounding word this could be a city or a person

ii) because as, we've spotted in example 2 above, the ~~addi~~ word itself could be polysemic and hence we need the context to decide

iii) More features which may help:

- Capitalisation (ie: a person is likely to start with a capital letter)
- its frequency (ie: a person is assume to be more rare than other words)
- its context words

b) i) $L \in \mathbb{R}^{|V| \times D}$; $h^{(t)} \in \mathbb{R}^{1 \times H}$; $\hat{y}^{(t)} \in \mathbb{R}^{1 \times C}$ $\Rightarrow e^{(t)} \in \mathbb{R}^{1 \times (2W+1) \cdot D}$; $W \in \mathbb{R}^{(2W+1) \cdot D \times H}$; $U \in \mathbb{R}^{H \times C}$

~~$\Rightarrow e^{(t)} \in \mathbb{R}^{1 \times (2W+1) \cdot D}$; $W \in \mathbb{R}^{(1+2W)D \times H}$; $U \in \mathbb{R}^{H \times C}$~~ ;

$b_1 \in \mathbb{R}^{1 \times H}$; $b_2 \in \mathbb{R}^{1 \times C}$

ii) $e^{(t)} \longrightarrow O((2W+1)D)$

$h^{(t)} \longrightarrow O((2W+1)D \cdot H + H) = \overset{\text{(big O)}}{O}((2W+1)DH)$

$\hat{y}^{(t)} \longrightarrow O(H \cdot C + C) = O(HC)$

$\text{Cost Per Word} = O(\text{Cost}(e^{(t)}) + \text{Cost}(h^{(t)}) + \text{Cost}(\hat{y}^{(t)}))$

$= O((2W+1)D + (2W+1)DH + HC)$

$= O((2W+1)DH + HC)$

$\Rightarrow$ Cost Sequence of T words $= O((2W+1)DHT + HC)$

Scanned by CamScanner

d) i) best-token-level F1-score is with PER : person.
Its confusion matrix shows that it is more likely to be misclassified as an Organization than any other class

ii) One issue is that the predictions are independent of each other

2) NER with RNN

a) i) 1 more parameter : W_h

ii) $e^{(t)} \longrightarrow O(D)$

$h^{(t)} \longrightarrow O(H \times H + DH + H) = O(H^2 + DH)$

$y^{(t)} \rightarrow O(HC + C) = O(HC)$

$\Rightarrow$ Cost-for-T-words $= O(TH^2 + THC + TDH)$

b) i)

ii) because F_1 is not differentiable

d) i) Without the masking vector ~~with~~ $m^{(t)}$, we would have taken into account words for which we set the features to zero by cropping long sequence $\Rightarrow$ our loss will be "artificially" very low $\Rightarrow$ gradients will be updated incorrectly

3) NER with GRU

a) i) ~~$U_h = 1$ $W_h = 0$ $b_h = 0$~~

~~ii) $W_r = U_r = b_r = b_z = b_h = 0$~~ | ~~$W_z = 0$; $U_z = 1$; $W_h = 0$; $U_h = 0$~~

b)