CS224N Assignment 3

Kangwei Ling

June 16, 2017

1 A window into NER

(a) (i.)
- **Papa Johns** make the best pizzas in America.
- **The Goldman Sachs** are the leading global investment banking, securities and investment management firm.

(ii.) The word itself might be ambiguous and it may convey different meaning. Thus using features, gives an overall meaning.

(iii.)
- Part of speech (POS) tags
- Context words

(b) (i.)

$$\begin{aligned} \boldsymbol{e}^{(t)} &: (2w+1)D \\ \boldsymbol{W} &: (2w+1)DH \\ \boldsymbol{U} &: HC \end{aligned}$$

(ii.)

$$T \cdot [(2w+1)D + (2w+1)DH + HC + C]$$

(c) `q1_windoq.py`

(d) i. BEST F_1 score:

	P	R	F_1
Entity-level	0.81	0.85	0.83

Confusion Matrix

Table 1: My caption

go\gu	Per	Org	Loc	Misc	0
PER	2937.00	60.00	81.00	17.00	54.00
ORG	129.00	1671.00	117.00	57.00	118.00
LOC	39.00	107.00	1861.00	32.00	55.00
MISC	35.00	72.00	43.00	1011.00	107.00
O	38.00	55.00	19.00	30.00	42617.00

Mostly the model makes mistakes by recognizing PER as LOC, ORG as PER, LOC as ORG.

ii. The training data is skewed, as most of the words are O.

- Misclassification of ORG as PER

  ```
  x : Papa Johns make the best pizzas in America.
  y*: ORG  ORG   O O  O O O     LOC
  y’: PER  PER   O    O   O    O      O  LOC
  ```

- Misclassification of LOC as ORG

  ```
  x : New York city.
  y*: LOC  LOC  O
  y’: ORG LOC  O
  ```

2 Recurrent neural nets for NER

(a) i. Rnn has an extra parameter of H^2 for W_h and a parameter of $(2W + 1)$ less for W_x

ii. $T(VD + H^2 + DH + 2H + HC + C)$

(b)

(c) it is hard to directly optimze for F 1 because it requires predictions from the entire corpus to compute, making it very difficult to batch and parallelize.

(d) `q2_rnn.py`

(e) i. Without masking, the loss and gradient of the model would be evaluated on many non existential data. The gradients from the padding input would flow through the hidden state and affect the learning of the parameters.

ii. `q2_rnn.py`

(f) `q2_rnn.py`

(g) `q2_rnn.py`