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	O
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
O	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*: O   O  O ORG  ORG ORG     ORG ORG O O  LOC     O
    y': O   O  O ORG  O   LOC     O   ORG O O  LOC     O
    ```

    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 LOG, 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*: O      O MISC MISC MISC O             O       O
    y': O      O LOC  MISC MISC O             O       O
    ```