CS224N: NATURAL LANGUAGE PROCESSING WITH DEEP LEARNING ASSIGNMENT #3

ANTHONY HO

1. (a) (i) Example 1: "Stanford is great." - where "Stanford" could refer to Stanford University (organization) or a person with last name Stanford (person).
    Example 2: "I am going to Stanford." - where "Stanford could refer to Stanford Unviersity (organization) or Stanford, California (location).

    (ii) Because the features apart from the word itself could provide additional information and context not contained in the word itself which might help with removing ambiguity in identifying its named entity.

    (iii) Feature 1: the adjacent words from the word itself could be helpful in predicting whether the word is part of a named entity or not. For example, if the word is immediately succeeded by an action verb, it makes the word more likely to be a named entity.
    Feature 2: capitalization of the word could also be helpful in predicting whether the word is part of a named entity or not, especially in the case of person, organization, and location.

    (b) (i) The dimensions are:

    $$\boldsymbol{e}^{(t)} \in \mathbb{R}^{1 \times D(2w+1)}$$

    $$W \in \mathbb{R}^{D(2w+1) \times H}$$

    $$U \in \mathbb{R}^{H \times C}$$

    (ii) The computational complexity of predicting labels for a single word is $\mathcal{O}(D(2w+1) + D(2w+1)H + HC) = \mathcal{O}(D(2w+1)(H+1) + HC)$. Therefore the computational complexity of predicting labels for predicting labels for a sentence of length T is $\mathcal{O}(T(D(2w+1)(H+1) + HC))$.

    (c) Please see the coding portion of the assignment.

    (d) (i) The best development entity-level F_1 score is 0.83 and the corresponding token-level confusion matrix is shown below:

	PER	ORG	LOC	MISC	O
PER	2958.00	51.00	61.00	10.00	69.00
ORG	139.00	1659.00	120.00	47.00	127.00
LOC	48.00	148.00	1843.00	19.00	36.00
MISC	44.00	71.00	46.00	995.00	112.00
O	49.00	50.00	17.00	30.00	42613.00

    From the confusion matrix, it looks like the model has a tendency to misclassify organization as person, location or null, to misclassify location as organization, and to misclassify miscellaneous as null.

    (ii) (1) A window-based model would have troubles identifying named entities longer than the window itself. For example, the prediction made by our model (`window_size` = 1) on the sentence "The Senate Select Committee on Intelligence is investigating the Russian affairs ." is "O ORG ORG ORG O ORG O O O MISC O O", which fails to identify the "Senate Select Committee on Intelligence" as a single named entity instead of two.
    (2) A window-based model would fail to take long-range information into account, since it's based on a finite length window. For example, the prediction made by our model (`window_size` = 1) on the sentence "Washington was the first President of the United States ." is "LOC O O O O O O LOC LOC O", which fails to take into the account of the word "President" that estabishes "Washington" as a person instead of a location.

2. (a) (i) This particular RNN model does not necessarily have more parameters in comparison to the window-based model, since the window-based model has a bigger W (due to a bigger window) than the RNN's

model's W_e, while the RNN model has the additional parameter W_h. The difference in number of parameters between the RNN model over the window-based model is $H^2 - 2wDH$.

(ii) The computational complexity mainly depends on the sizes of the matrix multiplications and scales linearly to the number of time steps. Thus, the computational complexity of predicting labels for a sentence of length T with the RNN model is $\mathcal{O}(T(H^2 + DH + CH))$.

(b) (i) Since the cross-entropy cost is computed at the token level, its minimization does not guarantee the minimization of the entity-level F_1 score, such as when there are multi-word entities in the training set.

(ii) F_1 score is not additive over training samples like cross-entropy cost, which makes it difficult to be used in stochastic gradient descent during optimization.

(c) Please see the coding portion of the assignment.

(d) (i)

(ii) Please see the coding portion of the assignment.

(e)

(f)

(g)

3. (a) (i)

(ii)

(b) (i)

(ii)

(c) Please see the coding portion of the assignment.

(d)

(e)

(f)