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	2899.00	51.00	88.00	22.00	89.00
ORG	125.00	1668.00	95.00	72.00	132.00
LOC	31.00	112.00	1867.00	30.00	54.00
MISC	31.00	43.00	41.00	1037.00	116.00
O	33.00	42.00	18.00	34.00	42632.00

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

(b)

(c)

(d)

(e)

(f)

(g)

3. (a)

(b)

(c)

(d)

(e)

(f)