Question 1, Assignment 5, CS224n

(a) Typically, the vocabulary size of characters is lower than that of words, so a lower dimension of a lookup table is enough for character-embeddings.

(b) $Size_{char} = e_{char} * V_{char}$ and $Size_{word} = e_{word} * V_{word}$, the second has more parameters by a $\frac{50,000}{96}$.

(c) Since CNN is not sensible to the word positions, it's more robust to noises in the sentence, while RNN relies a lot on the positional information.

(d) Average-pooling makes more use of the information than max-pooling as it aggregates the global information by averaging, while max-pooling only focuses on max-informative part.

Finally the trained NMT System has 395.37 ppl over dev set. The BLEU over testset is 24.43.

(a) *traducir* and *traduce* occur. For word-based NMT, though these words have almost the same meaning, the occurance count may not count in some of them and ¡unk¿ problem will happen because of less appearance of some forms. However, char-aware NMT model can repair the issue because their forms are very similar.

(b) (i) Word based:

i. financial: economic, business, markets, banking, finance
ii. neuron: nerve, neural, cells, brain, nervous
iii. Francisco: san, jose, diego, antonio, california
iv. naturally: occurring, readily, humans, arise, easily
v. expectation: norms, assumptions, policies, inflation, confidence

(ii) Char based:

i. financial: vertical, informal, physical, cultural, electrical
ii. neuron: neurons, neurons., neuro, neuron,
iii. Francisco: Francisco, Francisco,, Francisco.
iv. naturally: pratically, typically, significantly, mentally, gradually
v. expectation: expectations, expectation

(iii) Word based embeddings focuse on the meaning of the word while char based embeddings focuse form. So in the results, the closest words for word based embeddings are more closer in terms of meanings than forms. By constrast, the closest words for char based embeddings are more closer in terms of forms instead of meanings.

(c) (i) Incorrect Example
Spanish: Puedo vestirme como agricultor, o con ropa de cuero, y nunca nadie ha elegido un agricultor.
Gold: You can have me as a farmer, or in leathers, and no one has ever chose farmer.
Char-based: I can dress as a farmer, or with leather clothes, and never nobody has chosen a farmer.
Word-based: I can <unk>like <unk>or with clothing and never chosen a farmer.
Possible Explanation: The multi forms of the word *vestirme* make the model difficult to learn the pontential meaning 'get dressed' and further predict it.

(ii) Correct Example
Spanish: Bien, al da siguiente estbamos en Cleveland.
Gold: Well, the next day we were in Cleveland.
Char-based: Well, the next day we were in Cleveland.
Word-based: Well, the next day we were in ¡unk¿.
Possible Explanation: The low frequency of *Cleveland* in the vocabulary makes the model difficult to learn effective information about the word and further predict it.