Assignment 5

1 a) In Assignment 4 we used 256-dimensional word embeddings ($e_{word} = 256$), while in this assignment, it turns out that a character embedding size of 50 suffices ($e_{char} = 50$). In 1-2 sentences, explain one reason why the embedding size used for character-level embeddings is typically lower than that used for word embeddings.

Ans) The embedding size for characters is smaller than the embedding size for words because the character vocabulary is much smaller than the word vocabulary, therefore a smaller embedding size is sufficient to represent each character.

b) Write down the total number of parameters in the character-based embedding model (Figure 2), and then do the same for the word-based lookup embedding model (Figure 1). Write each answer as a single expression (though you may show working) in terms of e_{char}, k, e_{word}, V_{word} (the size of the word-vocabulary in the lookup embedding model) and V_{char} (the size of the character-vocabulary in the character-based embedding model). Given that in our code, k = 5, $V_{word} \approx 50,000$ and $V_{char} = 96$, state which model has more parameters, and by what factor (e.g. twice as many? a thousand times as many?).

Ans) For character-based embedding model

No. of parameters $= e_{char}*V_{char} + f*e_{char}*k + f + e_{word}*e_{word} + e_{word} + e_{word}*e_{word} + e_{word}$

$= 50*96 + 256*50*5 + 256 + 256*256 + 256 + 256*256 +256$

$= 4800 + 64000 + 256 + 65536 +256 + 65536 + 256$

$= 200640$

For word-based embedding model

No. of parameters $= e_{word}*V_{word}$

$= 256*50000$

$= 12800000$

The word embedding model has more parameters by factor of approximately 64.

c) In step 3 of the character-based embedding model, instead of using a 1D convnet, we could have used a RNN instead (e.g. feed the sequence of characters into a bidirectional LSTM and combine the hidden states using max-pooling). Explain one advantage of using a convolutional architecture rather than a recurrent architecture for this purpose, making it clear how the two contrast. Below is an example answer; you should give a similar level of detail and choose a different advantage.

When a 1D convnet computes features for a given window of the input, those features depend on the window only – not any other inputs to the left or right. By contrast, a RNN needs to compute the hidden states sequentially, from left to right (and also right to left, if the RNN is bidirectional). Therefore, unlike a RNN, a convnet's features can be computed in parallel, which means that convnets are generally faster, especially for long sequences.

Ans) A convnet allows us to compute several local features of the characters in a particular window size (i.e. the size of the filter) while a RNN computes hidden states sequentially, hence it might not be able to capture only local features, because in theory even the first character can influence the last character. For some applications we might require these local features which can be computed efficiently and reliably by a convnet.

d) In lectures we learned about both max-pooling and average-pooling. For each pooling method, please explain one advantage in comparison to the other pooling method. For each advantage, make it clear how the two contrast, and write to a similar level of detail as in the example given in the previous question.

Ans) Max pooling takes the prominent i.e. the highest activation. The advantage is that the max value will sometimes be best value to describe its region. The disadvantage is that it discards the other values. In contrast average pooling takes the average of all the values, thereby preserving the information from the values of the region. The disadvantage is that because average pooling does not produce the most prominent signal it is sensitive to outliers and may not always produce the best word representations.

h) Highway Network: Once you've finished testing your module, write a short description of the tests you carried out, and why you believe they are sufficient.

Ans) The first test I performed was to make sure that all the tensors had the correct shape. The next test I performed was given an input tensor do the output an all the intermediate values match the ones calculated manually. A subtest of this test was to initialize the input as tensor of zeros. I think these tests are sufficient because they care of all the different cases that might be encountered at run time.

i) CNN: Once you've finished testing your module, write a short description of the tests you carried out, and why you believe they are sufficient.

Ans) The first test I performed was to make sure that all the tensors had the correct shape. The next test I performed was given an input tensor do the output an all the intermediate values match the ones calculated manually. A subtest of this test was to initialize the input as tensor of zeros. I think these tests are sufficient because they care of all the different cases that might be encountered at run time.

2 f) Report your test set BLEU score.

Ans) BLEU score = 23.737

3 a) State which of the six forms of the verb 'traducir' occur and which do not. Explain in one sentence why this is a bad thing for word based NMT from Spanish to English. Then explain in detail (approximately two sentences) how our new character-aware NMT model may overcome this problem.

Ans) The forms of the verb that occur are 'traducir' and 'traduce'. The words that do not occur are 'traduzco', 'traduces', 'traduzca' and 'traduzcas'. Since only some forms of the verb 'traducir' occur in the vocabulary whenever the model sees a word that is not part of the vocabulary it will map it to

the unknown word token and therefore the output will have an unknown word token. The new character-aware NMT model may overcome this because when a form of the verb that does not appear in the vocab is mapped to the unknown word token, the character model can reconstruct the word character by character and therefore the will not contain an unknown word token but will contain the correct form of the verb.

b) i) Go to https://projector.tensorflow.org/. The website by default shows data from Word2Vec. Look at the nearest neighbours of the following words (in cosine distance).

- financial
- neuron
- Francisco
- naturally
- expectation

For each word, report the single closest neighbour.

Ans) financial – economic (0.337)

neuron – neurons (0.439)

Francisco – san (0.171)

naturally – occurring (0.447)

expectation – operator (0.552)

ii) Load the character based word embeddings and again report the closest neighbour of each of the words given above.

Ans) financial – vertical (0.301)

neuron – Newton (0.354)

Francisco – France (0.420)

naturally – practically (0.302)

expectation – exception (0.389)

iii) Compare the closest neighbours found by the two methods. Briefly describe what kind of similarity is modelled by Word2Vec. Briefly describe what kind of similarity is modelled by the CharCNN. Explain in detail (2-3 sentences) how the difference in the methodology of Word2Vec and a CharCNN explain the differences you have found.

Ans) Word2Vec models the similarity between words i.e. it models the semantic relationship of words. CharCNN models the similarity between words by the similarity of the character they are made up of. Word2Vec represents each word by a unique word vector whereas a CharCNN represents each character by a unique vector, and uses these character embeddings to compose the

meaning of a word. Therefore the CharCNN models similarity of characters while Word2Vec models similarity of words.

c) Find places where the word based model produced <UNK> and compare to what the character based decoder did. Find one example where the character based decoder produced an acceptable translation in place of <UNK>, and one example where the character based decoder produced an incorrect translation in place of <UNK>. For each of the two examples you should:

1. Write the source sentence in Spanish. The source sentences are in en_es_data/test.es.
2. Write the reference English translation of the sentence. The reference translations are in en_es_data/test.en.
3. Write the English translation generated by the model from Assignment 4. Theses translations are in outputs/test_outputs_a4.txt. Underline the <UNK> you are talking about.
4. Write your character-based model's English translation. These translations are in outputs/test_outputs.txt. Underline CharDecoder generated word you are talking about.
5. Indicate whether this is an acceptable or incorrect example. Give a brief possible explanation (one sentence) for why the character based model performed this way.

Ans) Source Sentence: Un amigo mo hizo eso -- Richard Bollingbroke.

Reference Translation: A friend of mine did that -- Richard Bollingbroke.

A4 Translation: A friend of mine did that -- Richard <unk>

Character based model translation: A friend of mine did that -- Richard Bourner.

This is incorrect. The model could not predict the correct name may be because the unknown word embedding was not enough to generate the name 'Bollingbroke', it was only able to generate the first two letters of the word correctly.

Source Sentence: Una mujer autista llamada Zosia Zaks dijo una vez: "Necesitamos todas las manos en cubierta para enderezar el barco de la humanidad".

Reference Translation: An autistic [man] named Zosia Zaks once said, "We need all hands on deck to right the ship of humanity."

A4 Translation: A autistic woman named <unk> <unk> once said, "We need all the hands on cover for <unk> the <unk> <unk>

Character based model translation: An autistic woman named Zolan Zappies once said, "We need all the hands in covered the book of humanity.

This is correct. The model is correctly able to generate the word humanity from the unknown vector embedding.