Assignment 4

CPSC 599/601: Practical Natural Language Processing

March 13, 2024

This Assignment will evaluate your skills in information extraction, and text generation. Include in your submission the code used to generate answers as a Jupyter Notebook or Python program file as well as any files generated. Make sure it is clear what code answers each question.

This Assignment is meant to be completed individually. You may discuss the questions at a high level with other students but the final work submitted must be your own. Please reference any external resources you use to complete this Assignment using ACM referencing format.

Exercise 1 (15 marks)

A) Describe a set of 5 potential features for a sequence-labeling classification using CRF. (**5 marks**)

B) Explain why these features tend to be useful for these types of sequence labeling tasks. (**10 marks**)

Ex2: you have provided the code for, provide the trace.

Exercise 2 (20 marks)

You have been provided a basic LSTM for text generation in the notebook provided for this assignment. Use it to answer the questions below.

A) Trace the steps through the LSTM provided and explain in 1-2 sentences what each line is doing (similar to how RNN is commented in the text generation notebook in class). (**10 marks**)

B) Explain how temperature is expected to impact text generation. Show examples with generated text with temperature set to 0.1, 0.5, 1, and 1.5. Do you see any evidence of temperature impacting the results? (**8 marks**)

Any interesting things

C) Describe 2 benefits a LSTM has over a vanilla RNN. Describe 1 limitation of an LSTM. (**2 marks**)

Exercise 3 (15 marks)

A) Please submit your generative chatbot made following the in-class activity alongside an example set of dialog between yourself and the chatbot before and after fine-tuning on a character of your choice. It is recommended to use a larger dataset than the one in class for better performance. Describe your results including what happens when you increase the number of epochs to train the model and if there were any limitations with your selected dialogue dataset? (**14 marks**)

B) What is the maximum sequence length that DialoGPT can accept by default? Hint: the max length is used in the code during conversation generation so you can see how to access it. (**1 marks**)

Below is an example of DialoGPT before and after fine-tuning with the in-class activity.

Today exercise notebook

How to prepare dataset to run DialoGPT (fine-tune of GPT 2)

```
No fine-tuning:
>> User: Hi.
DialoGPT: I'm not your buddy, friend!
>> User: Okay. Who are you then?
DialoGPT: I'm not your buddy, pal.
>> User: That's fine.
DialoGPT: I'm not your guy, buddy.
>> User: Are you Canadian?
DialoGPT: I'm not your guy, guy.
>> User: Well, I think that's enough.
DialoGPT: I'm not your buddy, guy.

Fine-tuning with 8 epochs:
>> User:Hi.
MarlinBot:  Hey son. Daddy's here. Hey! Hey, son. Hey, wait a minute.
>> User:Why?
MarlinBot:  Because you touch the tops of my tentacles, son!
>> User:Where are you going?
MarlinBot:  Oh, thank goodness.
>> User:What?
MarlinBot:  That's my boy.
>> User:Marlin?
MarlinBot: Nemo!! I'm not gonna lose you again!! Okay, I'm done.

* Note that this was the most coherent response I was able to get.
Responses will vary after fine-tuning and depend on the amount of training data used.
Changing the temperature, top_p, and top_k settings will also impact the quality
of the responses. You may also experience responses of a lot of punctuation with the
small version of DialoGPT.
```