CMPT 413/825 Natural language processing
Homework 4, Fall 2020

Due: November 12th, 2020

Instructions Answers are to be submitted by 11:59pm of the due date as a PDF file. Please submit your answers using either Gradescope (preferred) or Coursys. You do not need to submit your answers using both platforms, just one.

- **Gradescope**: Go to https://www.gradescope.ca/, select **HW4** and submit your PDF. Please make sure to assign the pages to corresponding questions.
- **Coursys**: Go to Coursys, select the **Homework 4** activity and submit your answer as **answer.pdf**.

This assignment is to be done individually.

1. Analyzing NMT Errors (12 points)

Here we present a series of errors we found in the outputs [1] of our NMT model (which is what you will implement in HW4-P). For each example of a Spanish source sentence, reference (i.e., 'gold') English translation, and NMT (i.e., 'model') English translation, please:

1. Identify the error in the provided **NMT translation**.
2. Provide possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
3. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

Below are the translations that you should analyze as described above. Note that out-of-vocabulary words are underlined. Rest assured that you don't need to know Spanish to answer these questions. You just need to know English! The Spanish words in these questions are similar enough to English that you can mostly see the alignments. If you are uncertain about some words, please feel free to use resources like Google Translate to look them up.

(a) (2 points) **Source Sentence**: *Aquí otro de mis favoritos, "La noche estrellada".*
Reference Translation: *So another one of my favorites, "The Starry Night".*
NMT Translation: *Here's another favorite of my favorites, "The Starry Night".*

(b) (2 points) **Source Sentence**: *Ustedes saben que lo que yo hago es escribir para los niños, y, de hecho, probablemente soy el autor para niños, ms ledo en los EEUU.*
Reference Translation: *You know, what I do is write for children, and I'm probably America's most widely read children's author, in fact.*
NMT Translation: *You know what I do is write for children, and in fact, I'm probably the author for children, more reading in the U.S.*

(c) (2 points) **Source Sentence:** *Un amigo me hizo eso – Richard Bolingbroke.*
Reference Translation: *A friend of mine did that – Richard Bolingbroke.*
NMT Translation: *A friend of mine did that – Richard <unk>*

[1] The data is from TED talks.

(d) (2 points) **Source Sentence:** *Solo tienes que dar vuelta a la manzana para verlo como una epifanía.*
Reference Translation: *You've just got to go around the block to see it as an epiphany.*
NMT Translation: *You just have to go back to the apple to see it as an epiphany.*

(e) (2 points) **Source Sentence:** *Ella salvó mi vida al permitirme entrar al baño de la sala de profesores.*
Reference Translation: *She saved my life by letting me go to the bathroom in the teachers' lounge.*
NMT Translation: *She saved my life by letting me go to the bathroom in the women's room.*

(f) (2 points) **Source Sentence:** *Eso es más de 100,000 hectáreas.*
Reference Translation: *That's more than 250 thousand acres.*
NMT Translation: *That's over 100,000 acres.*

2. Contextualized Word Embeddings (14 points)

(a) (3 points) Some word embeddings, such as ELMo[2], use character level embedding rather than word level embedding. What might be an advantage of using character level embedding over word level embedding.

(b) (4 points) Another word embedding called BERT[3] uses masked language modeling as its pre-training objective. Explain what this objective is. Also mention an advantage of using this objective in some downstream natural language tasks.

(c) (4 points) Your friend wants to build an entity linking system using BERT. In an entity linking system, named entity references in the text are linked to their corresponding entities in a knowledge base.

For example, consider the following text:
"*Vancouver* is one of the most ethnically and linguistically diverse cities in Canada..."

Then the named entity *Vancouver* can be linked to the corresponding page for the city *Vancouver* from DBpedia[4] *http://dbpedia.org/page/Vancouver*

Assume that you have a training dataset with sentences and spans of tokens linked to entities in DBpedia. You want to help your friend build the entity linking system. To start, you will need to design a model that, given text as input, will need to identify what spans need to be linked and what the target entities should be for each span.

Explain how you would setup an entity linking model using BERT. During training, what input would you provide to the network? What would the output of the network be? How would you train the network (what parameters will be learned or fine-tuned, and what would be the training objective at a high level)?

(d) (3 points) What are some advantages of the transformer model over RNNs? List at least two.

3. Constituency Parsing (12 points)

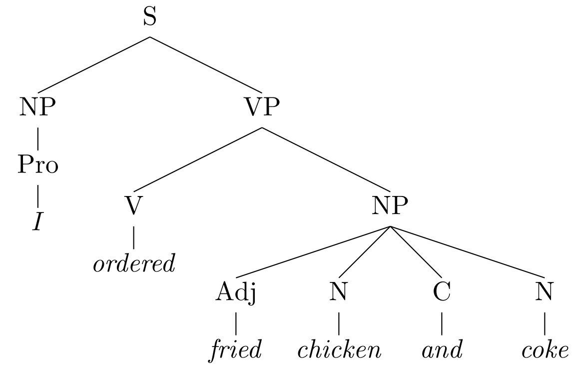
Consider the following treebank consisting of three sentences and their parse trees. The part of speech tags are D (determiner), Adj (adjective), N (noun), V (verb), P (preposition), C (conjunction), Pro (pronoun) and the phrases are NP (noun phrase), VP (verb phrase), PP (prepositional phrase) and S (sentence).

1. I ordered fried chicken and coke
2. It rains this time of the year
3. The little kids are playing violin at the concert

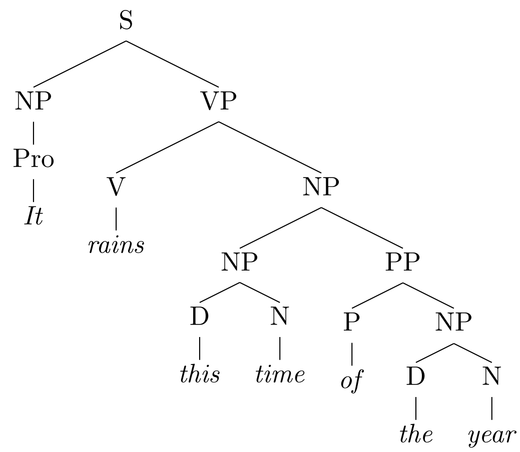
Constituency structure:

[2]Deep contextualized word representations: `https://arxiv.org/pdf/1802.05365.pdf`
[3]BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: `https://arxiv.org/pdf/1810.04805.pdf`
[4]DBpedia is a knowledge base extracted from Wikipedia

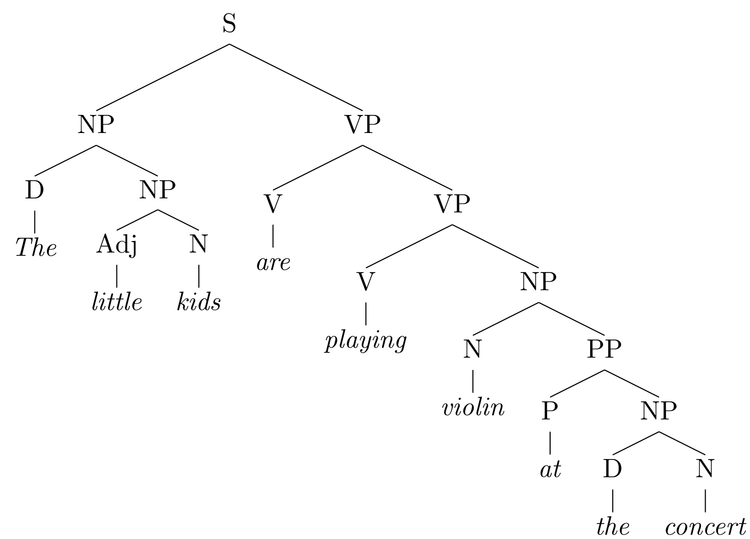
1.



2.



3.



(a) (2 points) We would like to learn a probabilistic constituency parser from the above trees. As a first step, we need to define the Probabilistic Context Free Grammar (PCFG). Specify what is the set of non-terminals, terminals, and start symbol in your PCFG.

(b) (2 points) Mention some production rules that the CFG would have. You do not need to enumerate all production rules, but make sure to include examples of rules involving *non-terminals-to-non-terminals* and *non-terminals-to-terminals*.

How many rules would we have for each of the types *non-terminals-to-non-terminals* and *non-terminals-to-terminals*, based on the three training examples that we have?

(c) (2 points) Specify what parameters you would need to learn for your PCFG. Describe how you can estimate the parameters.

(d) (4 points) For the first sentence in the above treebank, i.e. "I ordered fried chicken and coke", draw an alternate parse tree than the given one, which might give us an ambiguous meaning. Additionally, explain how you would handle disambiguation, i.e. how you would choose one parse tree over another for a single sentence.

(e) (2 points) What are some advantages of using neural networks over using statistical methods for constituency parsing? List at least two.

4. Transition-Based Dependency Parsing (12 points)

Let us consider the transition-based dependency parser (Figure-1) and three transition operators as below[5]:

```
function DEPENDENCYPARSE(words) returns dependency tree

  state ← {[root], [words], [] }  ; initial configuration
  while state not final
    t ← ORACLE(state)       ; choose a transition operator to apply
    state ← APPLY(t, state) ; apply it, creating a new state
  return state
```

Figure 1: Transition-based dependency parser

1. *LEFTARC*: Assert a head-dependent relation between the word at the top of the stack and the word directly beneath it; remove the lower word from the stack.
2. RIGHTARC: Assert a head-dependent relation between the second word on the stack and the word at the top; remove the word at the top of the stack.
3. SHIFT: Remove the word from the front of the input buffer and push it onto the stack.

(a) (2 points) What is the algorithmic complexity of the dependency parser in Figure-1?

(b) (6 points) Assuming an Oracle which provides correct transition operators, show the sequence of transitions for the following sentence:
He drove to the store to buy gifts

(c) (2 points) Mention two drawbacks of the dependency parser used for this question.

(d) (2 points) Explain how neural network can be used implement a transition based dependency parser. Specifically, what would the neural network be predicting and what would be inputs to the neural network.

[5]Algorithm for parser and the transition operator descriptions are taken from the Jurafsky book `https://web.stanford.edu/~jurafsky/slp3/15.pdf`. Please refer to the detailed example of the parse sequence in section:15.4 (Figure 15.7) in the book.