Assignment 4

CMSC 473/673 — Introduction to Natural Language Processing

Due Tuesday December 13th, 11:59 PM

Item	Summary
Assigned	Friday December 2nd
Due	Tuesday December 13th
Topic	Language Modeling and Insights
Points	80 (+30 EC)

In this assignment you will explore neural language models.

You are to *complete* this assignment on your own: that is, the code and writeup you submit must be entirely your own. However, you may discuss the assignment at a high level with other students or on the discussion board. Note at the top of your assignment who you discussed this with or what resources you used (beyond course staff, any course materials, or public Discord discussions).

The following table gives the overall point breakdown for this assignment.

Question	1	2	3	4
Points	25	25	30	30 EC only

What To Turn In Turn in a writeup in PDF format that answer the questions; turn in all requested code necessary to replicate your results. If you write code, be sure to include specific instructions on how to build and run your code. Answers to the following questions should be long-form.

How To Submit Submit the assignment on the submission site:

https://www.csee.umbc.edu/courses/undergraduate/473/f22/submit.

Be sure to select "`Assignment 4`."

Full Questions

1. (**25 points**) Consider a sequence on N input words $w_1 w_2 \dots w_N$ and a corresponding sequence of output labels $y = y_1 y_2 \dots y_N$. Assume we can represent each input word w_i via some embedding, e_{w_i}.

   In class, we looked at the following RNN cell definition, that processes the input sequence left-to-right:

   $$h_i = f(W h_{i-1} + U e_{w_i}), \tag{1}$$

   where f is a non-linearity that is applied component-wise to a vector (e.g., if $f = \exp$, then $f(1, 2) = (\exp(1), \exp(2))$). We would then form a distribution over our output labels as $\mathrm{softmax}(\theta h_i)$: this may often be written as $p(y_i | w_{\leq i}) = \mathrm{softmax}(\theta h_i)$, to indicate that distribution over output values for y_i is taking into account w_i and all previous words.

   In these equations, assume h_i is a K dimensional vector, each embedding vector e_v is an E dimensional vector, and there are L possible label types (that is, each y_i can be one of L possible values).

   (a) What does each of W, U, and θ serve to do?

   (b) What are the shapes of W, U, and θ?

   (c) If we were using this RNN to learn a language model, and we had the sentence "The can can hold liquid," what are each of the w_i and y_i values? (Hint #1: $N = 6$. Hint #2: Both w and y are of length 6 but they are not *exactly* equal. Hint #3: Remember that for language modeling you must learn to predict an EOS symbol.)

   (d) How would training an RNN as a language differ when using teacher forcing vs. not using teacher forcing?

2. (**25 points**) In class, we considered cases where these labels could be the original words themselves (so the RNN would learn a language model to predict what the next word is), part of speech tags, and named entity tags. One limitation of the RNN definition in the previous question is that only processes data left-to-right.

   (a) Give an example, either from part of speech tagging or named entity classification, where this left-to-right processing is a potential disadvantage. That is, give an example sentence w and labels y where you believe that predicting y_i would benefit from observing some word w_j that occurs after w_i.

   (Your example does not need to be an English example. But, if you provide an example in another language, you must provide an English explanation.)

   Luckily, there's an easy fix for this: it's called a bi-directional model, and all it does is it forms both a left-to-right representation l_i *and* a right-to-left representation r_i, and then concatenates them to form h_i. This changes the definition to:

   $$l_i = f(W l_{i-1} + U e_{w_i}) \tag{2}$$

   $$r_i = f(Z r_{i+1} + Q e_{w_i}) \tag{3}$$

   $$h_i = \mathrm{concat}(l_i, r_i) \tag{4}$$

   $$p(y_i | w_1, w_2, \dots, w_N) = \mathrm{softmax}(\theta h_i) \tag{5}$$

Here, we're introducing two additional parameter matrices (Z and Q).

(b) Why is this formulation okay to use for a problem like part of speech tagging or named entity recognition, but it would *not* be okay to use for a problem like language modeling?

(c) Consider the sequence $x =$ "cats eat" with part of speech label sequence $y =$ "Noun Verb." Under the following assumptions, compute each of $l_1, r_1, l_2, r_2, p(y_1|w_1, w_2), p(y_2|w_1, w_2)$, and each of the loss terms for y_1 and y_2. The assumptions are:

- The only two possible POS labels are Noun and Verb.
- Each word is represented as a 2-dimensional (column) vector: $e_{\text{cats}} = (0.75, 0.25)$ and $e_{\text{eat}} = (0.3, 0.2)$.
- $W = Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $U = Q = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$
- $\theta = \begin{pmatrix} \theta_{\text{Noun}} \\ \theta_{\text{Verb}} \end{pmatrix} = \begin{pmatrix} 0.7 & -0.4 & -0.3 & 1 \\ -0.2 & 0.35 & 0.4 & 1 \end{pmatrix}$
- The non-linearity $f = \tanh$.
- $l_0 = r_{N+1} = 1$ are both 2-dimensional vectors all of 1s.

If you would like, you may verify your computations using Pytorch. This is optional but if you do so, turn in your code. (Note that in this case, each of the matrix-vector products would be represented by a linear layer with `bias=False`.)

3. (**30 points**)

Read Bender et al. (2021). The article is available at `https://dl.acm.org/doi/10.1145/3442188.3445922`). After reading this article, think about the potential benefits and costs or harms of large language models (LLMs). Then, write a 1/2 page reflection piece. Your reflection should

- contain a brief summary of the findings/conclusions of this paper.
- explain your opinion on the benefits of LLMs, accounting for the potential costs.
- be understandable by a general UMBC student audience (don't assume someone has taken this class).

4. (**30 points EC**) In this question, you will put on a "probing" (linguist/science-based) hat. You will think about different types of syntactic and/or semantic phenomena, and see how well large langauge models (Transformer methods) can capture them, via notions of embedding similarity.

- **Your Task** Identify three different linguistic phenomena: call these P_1, P_2, P_3. For each phenomena P_i, construct a minimum of $M = 5$ pairs of sentences that address or demonstrate that phenomena. Use a pretrained, large language model (sometimes abbreviated LLM) to evaluate the similarity of each pair. Create and turn in a report in which you:
  - Provide the M sentences for each phenomena.

- Discuss how you are using evaluating the similarity of each pair.
- For each phenomena, present your similarity results.
- For each phenomna, summarize and analyze the results. Your analysis should discuss, based on this limited exploration, your take-aways on how well LLMs can capture these phenomena, and where they fall short.

Turn in all code.

- **How You Will Be Graded** Your grade on this problem will be on how well your report completes the above tasks.
- **Specifics** You may use a skeleton Colab notebook (`https://colab.research.google.com/drive/101XGwu8s9Ur9txzmipmGVcBI3qniWqZ6?usp=sharing`).

  The choice of which LLM to use is up to you, with one exception: you may **not** use either `bert-base-uncased` or `bert-base-cased`.

  The choice of how you evaluate similarity it also up to you.
- **Example** Let's say you were examining the following "a" vs. "an" phenomena:

  > "a" vs. "an": Generally, we use "a" before nominals that do not begin with a vowel (sound) and "an" before nominals that do. For example, we say "a banana," and "an apple," but we also say "a red apple."

  For this phenomena, you'd want to construct at least 5 pairs of phrases that focus on this syntactic rule. You want each pair to be appropriately contrastive. This contrast could be between (grammatically) acceptable and not [for example, "an apple" vs. "a apple"], or its interaction with other grammatical rules [e.g., "an apple" vs. "a red apple," or "an apple" vs. "an inedible apple"], or its interaction with spelling vs. sound [e.g., "an hour" vs. "a hour"], or other interactions *that you think of*. For each phrase, use a LLM to embed that phrase and compute how (dis)similar the LLM judges those phrases.

  I would *highly* recommend focusing one of these interactions and coming up your 5 pairs based on that single interaction. That is, don't try to evaluate grammatical acceptability *and* grammatical rule interaction *and* spelling vs. sound. Pick one and focus on that.

- **Possible Phenomena** The following are *some* possible phenomena. **You are not restricted to this list**, with the exception that you may not use the "a vs. an" phenomena (since it was given as an example). You may also **not use any examples given here**.

  Adjective Order There are many different ways that adjectives can modify nominals. Some modify the number ("two"), others modify a subjective value ("great"), and other modify physical/observable attributes (e.g., size, color and shape). With these modifiers, it is more natural to say "two great big green houses" than it is to say "great green two big houses."

  Transitive vs. Di-Transitive vs. Intransitive Verbs Certain verbs are transitive, meaning they can take direct objects ("Chris ran the marathon"); others are di-transitive, meaning they can take direct objects and indirect objects ("Chris gave Pat the book"); and others are intransitive, meaning they do not take any objects ("Chris ran toward the hills"). These may take certain modifiers, e.g., "Chris ran toward the hills with vigor," "Chris ran the marathon with vigor," and "Chris gave the book to Pat."

  Noun-Verb Agreement One of the most common instance of English's inflectional morphology is in noun-verb agreement. For example, "she loves NLP" vs. "they love NLP."

  Selection Preferences Certain verbs have preferences for certain types of arguments. For example, "drink" prefers its *subject* to be animate and its *object* to be drinkable, or at least consumable, as seen in the sentence "Chris drank the soda." Contrast this with the sentence "Chris drank the laptop," which just does not make sense. However, these aren't requirements: imagine getting poison ivy and putting lotion on the rash. You may say "the rash really drank the ointment." Similarly, some combinations just are more acceptable than others: "Chris drank the soda" is more acceptable than "Chris drank the gravy."

  Upward vs. Downward Entailment This concerns the inferences that can be made by using different types of quantifiers and determiners in sentences, and how they interact with different phrases in the sentence. For example, when we say "Some children like vegetables," and we focus on the interaction between the quantifier "some" and the noun "children," we can infer that "A child likes vegetables," but we cannot infer that "All children like vegetables." On the other hand, if we say "No child likes steamed vegetables," and we focus on the interaction of the determiner "no" and the phrase "steamed vegetables," we can infer that "no child likes steamed broccoli," but we cannot infer that "no child likes vegetables."

  Veridicality/factuality This concerns how completion of an event, or lack thereof, can be communicated. For example, "Pat ran a marathon" means that Pat actually ran the marathon (and completed it), while "Pat tried to run a marathon" does not necessarily mean that Pat actually completed it.

  Deverbal Events Many times we communicate an event via a verb, for example "Pat and Chris passionately argued over the plan." However, we can also communicate (similar, but not necessarily the same) information about an event through a nominalized form of that verb, such as "Pat and Chris's argument over the plan was

passionate," or "Pat and Chris's argument surprised people."

Subsective Adjectives This concerns how adjectives can interact with one another. In many cases, English adjectives are subsective: if we think of the noun being modified as a set, adjectives identify various subsets of that set, and they act intersectionally. For example, imagine the noun "cats" referring to a set of all possible cats; then "fluffy cats" is a subset of the "cat" set, and "white, fluffy cat" is a further (intersectional) subset. However, not all adjectives, such as "fake," are subsective. For example, "fake cats" would not refer to a subset of the "cats" set.