First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2022
Johns Hopkins University
Co-ordinator: Philipp Koehn

6 October 2022

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: ______________________________

Q1. Language Modeling *12 points*

Consider the sentence:

``After baking all day, I frosted and ate 3 very large w_{i+1}''

(1) (2 points) List two very likely candidates for w_{i+1} given a **factored language model**:

(a) ______________________________

(b) ______________________________

(2) (2 points) List two reasonably likely candidates for w_{i+1} given a **trigram language model**:

(a) ______________________________

(b) ______________________________

(3) (8 points) Give instantiations of $f(w_{i+1} \mid FEATURE,\ TYPE)$ that lead you to your factored LM candidates in (1) above, for **FOUR** specific instantiations of quite distinct feature types TYPE that are not simple adjacent n-grams:

(Example) $f(w_{i+1}|FEATURE = "large",\quad TYPE = PrecedingWord)$ ⇐ Don't use

(a) $f(w_{i+1}|FEATURE =$ __________________, $TYPE =$ ________________________________)

(b) $f(w_{i+1}|FEATURE =$ __________________, $TYPE =$ ________________________________)

(c) $f(w_{i+1}|FEATURE =$ __________________, $TYPE =$ ________________________________)

(d) $f(w_{i+1}|FEATURE =$ __________________, $TYPE =$ ________________________________)

The TYPEs above don't need to be formal or limited to a specific set or syntax. Please just try to make your meaning clear. Picking diverse feature types that can each contribute distinctly to a factored language model should be your main goal.

Q2. Morphology *8 points*

(1) Give an example of an English word that has multiple meanings due to morpheme bracketing ambiguities. (Please do NOT repeat a specific example given in the lecture):

WORD: ____

(2) give 2 different morpheme bracketings for the word, e.g. ”((build -er) -s)”:

BRACKETING #1: ____

BRACKETING #2: ____

(3) state/paraphrase what each of these word interpretations mean:

MEANING #1: ____

MEANING #2: ____

(4) Which is more likely to lead to bracketing ambiguities:
INFLECTIONAL MORPHEMES or **DERIVATIONAL MORPHEMES** (circle one).

Q3. Syntax *20 points*

Formal Language Theory

1. (2 points each) Give an informal description and example of each of the following concepts in formal language theory:

   - alphabet (Σ)

   - word (α, β)

   - language ($\mathcal{L}$)

   alphabet: a set of symbols (or tokens); word: a sequence of zero or more symbols from the alphabet (a "string"); language: a set of words (or strings)

2. (4 points) What is a natural language? Compare and contrast with synthetic languages, and the relationship of all of the to formal language theory.

   Natural languages are human languages, in contrast to synthetic or engineered languages or specifications like Python or HTML. They can both be split into groups of "valid" and "invalid" sentences, though the split is fuzzier in natural languages. Formal language theory can be used to describe the structure of both types of languages, though its application to natural languages is not perfect.

Linguistics

1. (2 points) Define a noun (or noun phrase) using the grammar school, distributional, or functional definition.

   grammar school (semantic definition): a person, place, thing or idea; *distributional*: nouns are the set of words and phrases that have the same sentence distributions; *functional*: the set of words/phrases that serve as arguments to verbs

2. (2 points) What is the Penn Treebank?

   Students should have some of the following facts: A collection of 50k sentences / 1 million words of data from the Wall Street Journal, parsed by human annotators, used to train models for parsing.

3. (2 points extra credit) List the first sentence of the Penn Treebank.

   Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

Parsing

1. (1 point) We discussed two reasons for parsing natural language. The first reason was to determine whether a sentence was in the language. Why is this question usually unhelpful to ask when using a grammar trained on the Penn Treebank?

   Grammars are very leaky, and can parse even really ungrammatical sentences, so the answer to the question "Is this sentence in the language"—which was central to formal language theory—is unfortunately almost always *yes*.

2. (1 point) What is the second purpose of parsing natural language?

   To find the structure of the sentence under some model. This is important for downstream tasks, including interpreting the sentence properly.

3. (4 points) Draw a parse tree above the following sentence. You should begin by labeling each word's part of speech (POS), and then connect these into spans hierarchically. Use the following POS inventory: determiner (D), adjective (A), noun (N), verb (V), preposition (P), and the following span labels: verb phrase (VP), noun phrase (NP), prepositional phrase (PP), and sentence (S).

   THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG.

   (S (NP (D the) (A quick) (A brown) (N fox)) (VP (V jumped) (PP over (NP (D the) (A lazy) (N dog)))))

Q4. Semantics *15 points*

1. (2 points) Which of the following is a *semantic role*?

    - a noun phrase (`NP`)
    - `INDIRECT_OBJECT`
    - the lexeme ”John”
    - `AGENT`

    `AGENT`

2. (13 points) Explain the purpose of the Lesk algorithm and why it doesn’t require annotated training data.

    The Lesk algorithm’s goal is word sense disambiguation (WSD), assigning the specific WordNet synset that a given word corresponds to in a particular context (sentence). Instead of annotating a new gold standard, it leverages the glosses (definitions) and examples provided by WordNet to compute their lexical overlap with the sentence being considered, selecting the most-overlapping synset as the likeliest assignment.

Q5. Deep Learning *20 points*

1. (5 points) Why is XOR an important problem for neural networks?

   Some sort of combination of: not linear separable and/or needs a hidden layer.

2. (5 points) Deep Learning has gotten more popular in recent years. Why is this the case? Please identify two (and only two) reasons.

   They need two from this list, or something that makes sense that I didn't think of. More digital data. Compute Power (GPUs), Availability of toolkits/Open Source focus

3. (5 points) Traditional Statistical Learning Algorithms frequently use features. How is this different than most Neural Network applications? Don't overthink this question.

   Basically fine with anything that says we do not use features as inputs often. Or, that they "learn" features.

4. (5 points) As an input to the embedding layer, how is a word (or sub-word unit) generally input into a neural network?

   1-hot vector. I'm inclined to give 4 points for "'look-up table' but I think full-credit is one-hot.

Q6. Information Retrieval *10 points*

1. (2 points) Suppose my Information Retrieval (IR) system returns 30 documents for a query. Among them, three documents are considered relevant. There are an additional seven relevant documents that are not returned. Please compute Precision and Recall.

2. (1 point) In practice, it may not be trivial to estimate the true Recall for web-scale IR systems that indexes millions of documents. On the other hand, computing Precision is possible with the help of some human annotators. Please explain why you think there may be such a difference.

3. (2 points) What is the difference between an information need and a search query? Please give an example of each.

4. (5 points) Suppose we have the following four sentences in our collection, and we want to build an IR system that retrieves sentences. Please draw a picture of an inverted index that indexes each sentence and enables fast retrieval.

   - $sentence_1$: Orioles beats Astros
   - $sentence_2$: Astros beats Orioles
   - $sentence_3$: Astros wins
   - $sentence_4$: Orioles wins

Q7. Information Extraction *25 points*

1. (10 points) Your goal is to build a system that collects daily weather information from weather reports in natural text (e.g., from radio broadcast transcripts) such as the following:

   > *Today Baltimore's weather will be sunny with near steady temperatures in the lower 70s. Northeast winds from 10 to 15 mph. Nights cool down to mid 50s.*

   What knowledge structure would you use for it? Sketch out its main elements.

   event frame: slots for location, date, temperature, wind speed and direction, amount of rain or snow, degree of cloud cover

2. (15 points) Assume that your knowledge base includes the following entities: *Baltimore Ravens*, *Baltimore Orioles*, *Baltimore City Government*.

   Now, you want to process the following text:

   > *Cortes, meanwhile, did not allow a hit in the first four innings, with walks to Jorge Mateo and Mountcastle accounting for* **Baltimore***'s lone base runners. Mateo's two-out single in the fifth ended the no-hit bid.*

   Given what you learned about neural models to represent meaning of words and text, how would go about building:

   - representations for the entities in knowledge base

     use BERT (or alike) to create embeddings for the entity from their occurrences in text

   - representations of the ambiguous word *Baltimore* in the text

     again, use BERT to obtain the word embedding

   - a method to use these representations to link the word in the text to the three candidate entities

     cosine distance or other similarity metric; train a classifier of examples of mention / entities pairs

Extra Space