First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2023
Johns Hopkins University
Co-ordinator: Philipp Koehn

5 October 2023

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Language Modeling *9 points*

1. In addition to word n-grams, list 3 other potential features appropriate for use in a factored language model. *(4 points)*

2. When building a LM for an automated telephone receptionist with 8 departmental options of equal probability (e.g. "say billing, scheduling, pharmacist, operator, etc.), what is the **upper bound** on this small LM's perplexity? *(5 points)*

   Generally describe a scenario that would substantially reduce that perplexity:

Q2. Morphology *11 points*

1. The following sentences A/B/C are translations into 3 artificial languages (of different major types) of the sentence: *(3 points)*

   I caused him to not visit Boston

   (A) VisktamehimBostonlermaiq .

   (B) Visktamehimlermaiq Boston .

   (C) Miya Boston tuu Nona maiqi Vesitanum .

   - Which is likely to be the Agglutinative language? A / B / C (circle one)
   - Which is likely to be the Isolating language? A / B / C (circle one)
   - Which is likely to be the Polysynthetic language? A / B / C (circle one)

2. Which type of language would likely be the **least** effectively handled by a learned morpheme segmentation model like Byte-Pair-Encoding: *(3 points)*

   (A) Agglutinative (e.g. Turkish)

   (B) Isolating (e.g. Chinese)

   (C) Templatic (e.g. Arabic)

   (D) Polysynthetic (e.g. Inuit)

3. Give a morphological segmentation and two derviational bracketings for the word "unwindable", along with the meanings of each (as a simple English meaning definition, in 3-6 English words each). *(5 points)*

   Segmentation and bracketing 1: _____

   Meaning 1: _____

   Segmentation and bracketing 2: _____

   Meaning 2: _____

Q3. Syntax *20 points*

Formal Language Theory

1. Formal language theory makes use of rules such A $\rightarrow$B C, A $\rightarrow$a, and B$\rightarrow$B c, where capital letters denote non-terminals symbols and lowercase letters denote terminal symbols. What is the difference between a terminal symbol and a non-terminal symbol? *(2 points)*

2. Draw a line between connecting each rule format in the Chomsky hierarchy with the associated language class. A, B, and C are nonterminals symbols, while α, β, and γ are strings of mixed nonterminal and terminal symbols. *(4 points, 1 for each correct link)*

rule format	**language class**
A$\rightarrow$B C	recursively enumerable
A$\rightarrow \alpha$	context-sensitive
αA$\beta \rightarrow \alpha\gamma\beta$	context-free
αA$\beta \rightarrow \gamma$	regular

3. Number the above language classes in terms of their descriptive power, with (1) being the most powerful, and (4) being the least powerful. *(4 points)*

4. What is the formal language definition of a language, $\mathcal{L}$? Give at least one example. (You can answer this formally or informally). What does it mean for a string to be *in* or *out* of a language? *(2 points)*

5. Explain the concept of "natural language". Given your definition, what are unnatural languages? Give an example of one. What is the use of formal language theory with respect to both types? *(2 points)*

6. Give an example of a grammar that recognizes the language of the domain portion of URLs, e.g., `www.ncats.net`, `https://microsoft.com`, and `jhu.edu`. Don't worry about the file path portion. *(6 points)*

Your non-terminal set should include the following nonterminals:

label	**description**
URL	The top-level symbol
SCHEME	http:// or https://
SUBDOMAIN	the subdomain (e.g., the *www* portion of `www.ncats.net`)
NAME	the main portion of the domain (e.g., jhu)
TOP	the top-level organizational domain

You can write the rules somewhat informally, e.g., using regular expression syntax (e.g., `[a-z0-9]+` for one or more letters or digits, `?` to mean one or more of the previous, and `*` to mean zero or more of the previous.). Do not worry about getting detail right, but seek to convince the grader that you understand the basic ideas.

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

Q4. Semantics *15 points*

1. What important distinction does WordNet *not* make between the different senses of a single word? Give an example of a word that exhibits this distinction. *(5 points)*

2. Explain why the Lesk algorithm, while sometimes called "unsupervised", would actually be very expensive to replicate for a newly-discovered language. *(10 points)*

Q5. Deep Learning *20 points*

1. What is supervised learning? *(5 points)*

2. Give 3 examples of supervised learning tasks in HLT. Show one input and output example for each. *(8 points)*

3. Name one thing that changed to make Deep Learning more popular now. *(7 points)*

Q6. Information Retrieval *10 points*

1. Suppose you are developing an IR system to retrieve breaking news about sports events. Your user enters the search query "orioles baseball" and the system retrieves the following three sentences. Sentence iii is undesirable because it is unrelated to the user's intent. First, please explain why this happens based on what you know about inverted indices. Second, suggest how you might modify your entire IR engine pipeline to prevent this error. *(6 points)*

   (a) Baltimore Orioles clinched AL East, capturing first division title since 2014

   (b) Orioles enters MLB playoffs as the most exciting storyline to watch

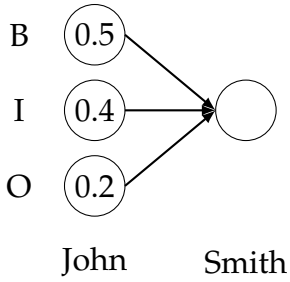
   (c) How to Attract Orioles and other Blackbirds to your Backyard

2. Embedding and Neural Net models for IR are gaining in popularity in research. Please describe the pros and cons of this approach in contrast to the traditional IR engine based on inverted indices. *(4 points)*

Q7. Information Extraction *20 points*

1. Viterbi Algorithm *(14 points)*

   Consider the following fragment of the search graph of the Viterbi algorithm.



$p(I|B) = 0.6$ $\quad$ $p(\text{Smith}|I) = 0.1$
$p(I|I) = 0.4$
$p(I|O) = 0$

   Using the Viterbi algorithm and the probabilities from the tables above, how what score and backpointer would the state on the right (corresponding to *(I; Smith)* assigned to?

2. What heuristics for pronoun interpretation would help you to disambiguate in the following cases: *(6 points)*

   - *John went to the ballgame with Jim. He bought beers and hotdogs.*
     Answer:

   - *Fritz is a big fan of Tottenham Hotspurs. He watched all the games. On Friday he went out with Paul. He asked the bar tender to put the game on the television.*
     Answer:

   - *Sue persuaded Ann that she should come to the party.*
     Answer:

Q8. Machine Translation *10 points*

1. What are BLEU scores? *(5 points)*

2. What is a problem with them? *(5 points)*

Extra Space