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)*

   part of speech ngrams, character ngrams, lemma ngrams, and any others that are plausible

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

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: _____

   [un [wind able]], not able to be wound. [[un wind] able], able to be unwound.

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)*

   A non-terminal symbol is one that is replaced by other terminal and non-terminal symbols in the string generation process, according to the rules provided in the grammar. A terminal symbol is one that does not undergo further replacement.

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

   The language classes are in reverse order, so the lines should reflect this.

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)*

   (1) recursively enumerable (2) context-sensitive (3) context-free (4) regular

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)*

   A language is a set of strings over a finite alphabet, Σ. It is a subset of the complete set of strings, Σ^*. Examples include the set of valid URLs, the set of floating point numbers, and the set of English-German code-switched sentences. A string s is not in a language if $s \in \Sigma^*$ and $s \notin \mathcal{L}$.

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)*

   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.

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:

   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.

   One solution is the following. The important part here is to capture the use of nonterminals in having responsibility for portions of the URL. For example, TOP captures the top-level domain (it doesn't list them all), and URL, the root symbol, captures how a domain comprises an optional scheme, an optional subdomain, the

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

domain, and the top-level domain. The SUBDOMAIN and DOMAIN comprise one or more characters followed by a period.

URL	→ SCHEME? SUBDOMAIN? DOMAIN TOP
SCHEME	→ https:// | http://
SUBDOMAIN	→ [a-zA-Z]+.
DOMAIN	→ [a-zA-Z]+.
TOP	→ `com` | `edu` | `net` | ...

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

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.. Full credit given if the sentence is more-or-less correct; partial credit is given if some of the major grammar or nouns are present in the sentence.

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)*

   WordNet does not distinguish between *polysemy* and *homonymy*, senses of the same word that are related or unrelated, respectively. Some examples are the different senses of the words ”lap”, ”fluke”, ”bank”, etc.

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

   The Lesk algorithm depends on vocabulary overlap between a token’s context and the definitions and examples provided by a lexical resource (e.g. a dictionary). Such lexical resources are quite labor-intensive, requiring manual compiling and annotation, so the approach is mainly useful for leveraging existing materials for high-resource languages.

Q5. Deep Learning *20 points*

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

   A type of machine learning training when there are correct labels.

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

   Any task like LangID, Speech Recognition

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

   Use best judgement. More data. GPUs. Standard Toolkits

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

   For the first part, students should demonstrate they understand how an inverted index looks like. They should describe having an dictionary entry "orioles" with postings that contain all the above sentences, which is why all are retrieved. For the second part, any creative and reasonable solution is ok. Examples include: use pseudo-relevance feedback exploiting the fact that "baseball" is also in the query, improve scoring function, add prior to sentences/documents from specific domains, do word-sense disambiguation, not including certain sites in document ingestion.

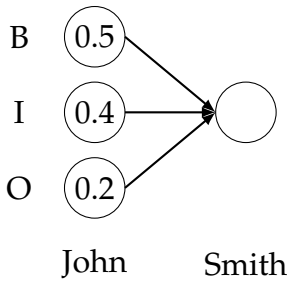
2. Embedding and Neural Net models for IR are gaining in popularity in research. Please describe the pros and cons of this approach in contrast to the traditional IR engine based on inverted indices. *(4 points)*

   Pros: (1) exact match is not required, so perhaps better handling of semantic similarity; (2) trainable so can exploit massive amounts of user data. Cons: (1) computationally more expensive to train and store. (2) slow at inference time, which affects responsiveness of applications. (3) labeled data might be insufficient. Students do not need to list all the above, but any response that demonstrates understanding of the tradeoffs is sufficient.

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?

   From B: $0.5 \times 0.6 \times 0.1 = 0.03$.
   From I: $0.4 \times 0.4 \times 0.1 = 0.016$.
   From O: $0.2 \times 0 \times 0.1 = 0$.
   Best is from B, with score 0.03.

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: Grammatical role; Parallelism.

   - *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: Repeated Mention.

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

Q8. Machine Translation *10 points*

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

   Automatic Metric for evaluation

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

   Best judgement. Synonyms. ETc.

Extra Space