First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2024
Johns Hopkins University
Co-ordinator: Philipp Koehn

3 October 2024

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Syntax *20 points*

Formal Language Theory

1. We defined a formal language as $\mathcal{L}$ as a set of words over an alphabet Σ, where the words are drawn from Σ^*.

   a. What is meant by Σ^*? *(1 point)*

   b. Give an example of a Σ and a language drawn from that set, such that $\mathcal{L}$ is a strict subset of Σ. *(1 point)*

   c. Suppose we define

$$\begin{aligned} \Sigma_1 &= \{0,1,2,3,4,5,6,7,8,9\} \\ \Sigma_2 &= \Sigma_1 \cup \{-\} \\ \Sigma_3 &= \Sigma_2 \cup \{.\} \end{aligned}$$

   What are three languages that could sensibly be defined over these alphabets?

   $\mathcal{L}_1$ =

   $\mathcal{L}_2$ =

   $\mathcal{L}_3$ = *(3 points)*

   d. For each language you have defined ($\mathcal{L}_\infty, \mathcal{L}_\in, \mathcal{L}_\ni$), give an example of a string that is *in* that language and a string that is *out* of the language. Recall that each string must be valid according to its respective alphabet. *(6 points)*

2. The Chomsky hierarchy defines four language classes of increasing power, corresponding to four contraints on the rule format. Draw a line between connecting each rule format in the Chomsky hierarchy with the associated language class. A, B, and C are nonterminal symbols, while α, β, and γ represent a mix of any number of nonterminal and terminal symbols (with a length of at least one). **Note**: one entry in each column is a fake. *(4 points, 1 for each correct link)*

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

3. Number the above language classes in terms of their descriptive power, with (1) being the most powerful, and (4) being the least powerful. *(2 points)*

4. 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? *(3 points)*

5. What is the first sentence of the Penn Treebank? *(1 point extra credit)*

Q2. Semantics *15 points*

1. Give an example of a word that is *polysemous* and a word that is *homonymous*, and explain why this is the case. *(10 points)*

2. Consider the sentences ″I baked the cake with the oven″ and ″The oven baked the cake″: characterize what syntactic and semantic roles the word ″oven″ has in each with respect to the verb ″baked″. *(5 points)*

Q3. Information Retrieval *10 points*

1. What is the difference between queries and topics? *(2 points)*

2. What information does relevance judgment capture in the Cranfield paradigm? *(3 points)*

3. Assuming you are designing a search engine for lay people (i.e., people without serious medical training) to search for medical literature with fast query response time, what kind of retrieval model architecture would you like to use and why? *(5 points)*

Q4. Distributional Semantics *15 points*

1. From what you know about embedding spaces for words that are learned with methods like CBOW, where would the following words be placed in relation to each other: *cat, lion, dog, friendly, dangerous*? *(10 points)*

   Put these words into the following box to reflect the distances of word vectors:

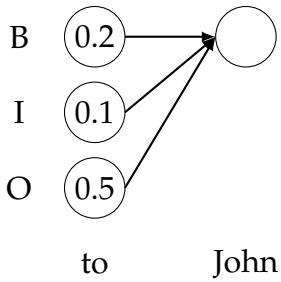
2. What is the difference between word embeddings (as learned by CBOW) and word representations in deeper layers of transformer models? Explain this for the example of the word *bat* in the following two sentences: *(5 points)*

   - *The player hit the ball with the bat out of the ballpark.*
   - *The explorer entered the cave where a bat flew just past his head.*

Q5. Information Extraction *20 points*

1. Viterbi Algorithm *(10 points)*

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



$p(B|B) = 0.1$ $p(\text{John}|B) = 0.1$
$p(B|I) = 0.2$
$p(B|O) = 0.3$

   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? Show all your computations that allow you to make that determination.

2. You were hired by a company to build a chatbot that provides information about the company. It should be able to answer questions like the following: *(10 points)*

   - What is Joe Johnson's cell phone number?
   - Can you give me the email address of Jane Smith in Accounting?
   - I would like to know in which department does Taylor Miller works. Do you have that information?

   Outline the design of such a chatbot.

Q6. Machine Translation *10 points*

1. Below is a bitext of 3 sentence pairs in two fictional languages, Alphian and Censian. Please list ALL the word translation equivalences based on code-breaking principles. *(5 points)*

   (a) Alphian: "SAN DOO" translates to Censian: "ce ti"

   (b) Alphian: "DOO MIO DET" translates to Censian: "ce bu ma"

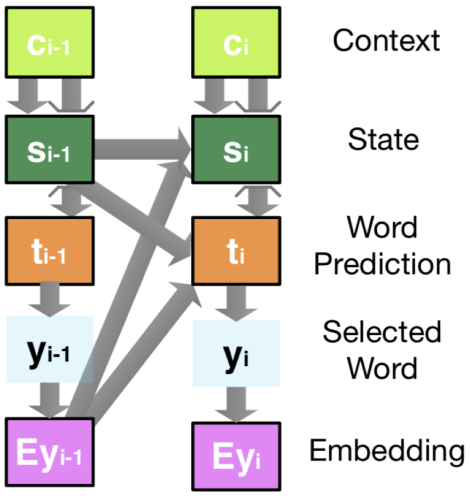
   (c) Alphian: "FAN MIO DOO" translates to Censian "ho bu ce"

2. Here is a schematic of an recurrent NMT decoder at the i-th step, where c_i represents the context vector coming from the encoder, s_i represents the recurrent hidden state, t_i represents the word prediction, and y_i represents the selected word. The state is defined by $s_i = f(s_{i-1}, Ey_{i-1}, c_i)$, where $f(\cdot)$ is a non-linear function. Please explain what might happen to the output translation if s_i is alternatively defined with fewer inputs. *(5 points)*

   (a) What happens if $s_i = f(Ey_{i-1}, c_i)$?

   (b) What happens if $s_i = f(s_{i-1}, c_i)$?

   (c) What happens if $s_i = f(s_{i-1}, Ey_{i-1})$?



Extra Space