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)*

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 $s \notin \mathcal{L}$.

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

$\alpha \rightarrow \beta$ and "context-laden" are fake. The rules are in order from regular, to context free, to context sensitive, to recursively enumerable.

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)*

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

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)*

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.

5. What is the first sentence of the Penn Treebank? *(1 point 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; one point of credit is given if the major gist is there.

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)* The word ”fluke” can be used for both polysemy and homonymy. For the former, it may refer to ”an unusual event” or ”a type of flat fish”, from unrelated roots from different languages (homonymy). For the latter, the meaning of ”a flat-shaped anchor” was coined due to the similarity with the fish (therefore, polysemy).

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)* In the first sentence, ”oven” is part of a prepositional phrase (i.e. the dependent of ”with”) attached to the verb, while in the second sentence it is the verb’s direct object. In both cases, its semantic role is ”instrument”, something non-intentionally performing an action.

Q3. Information Retrieval *10 points*

1. What is the difference between queries and topics? *(2 points)* A topic conveys the exact information that a user is looking for; a query is a realization of a search topic, which may be short, ambiguous, or even not present to the search engine (e.g., in a recommendation system where a user profile is an implicit query).

2. What information does relevance judgment capture in the Cranfield paradigm? *(3 points)* Relevance judgments record the accessor/annotator's opinion on whether a document is relevant to the topic. Such judgments are not factual but opinions since a document can be considered relevant for an accessor but not for others.

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)* Option 1 (a better one): learned-sparse retrieval model. Since the search engine needs to respond to queries quickly, learned-sparse retrieval models leverage the inverted index to provide short query latency while using a trained, neural sparse encoder to encode the queries and the documents. Such encoders provide semantic matching capabilities, which can bring the lay-person terms closer to the medical terms (or vice versa).
   Option 2: statistical retrieval model with dictionaries to match the terminologies between lay-person and medical literature. Statistical models leverage inverted index data structure for fast query latency. However, such models can only match the queries and documents based on their surface form. The system would need to expand or modify the queries or documents to ensure terms on either side are added to the queries or documents to ensure surface form matching is capable of retrieving relevant documents.

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:

   cat, lion, dog next to each other, with cat and dog closest; friendly and dangerous in a different cluster, friendly a bit closer to cat/dog, dangerous a bit closer to lion

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)*

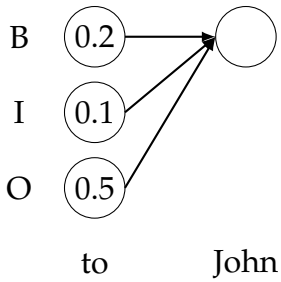
   - *The player hit the ball with the bat out of the ballpark.*
   - *The explorer entered the cave where a bat flew just past his head.*

   In CBOW word embeddings, the word *bat* would have the same fixed vector in both sentences, in deeper transformer layers they would be different, one closer to the representation of *racket*, one closer to the representation of *bird*.

Q5. Information Extraction *20 points*

1. Viterbi Algorithm *(10 points)*

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



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

   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.

   From B: $0.2 \times 0.1 \times 0.1 = 0.002$.
   From I: $0.1 \times 0.2 \times 0.1 = 0.002$.
   From O: $0.5 \times 0,3 \times 0.1 = 0.015$.
   Best is from O, with score 0.015.

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.

   expected answer: store information in a database, use a language model to translate the question into an SQL query, execute the query and return the answer.

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"

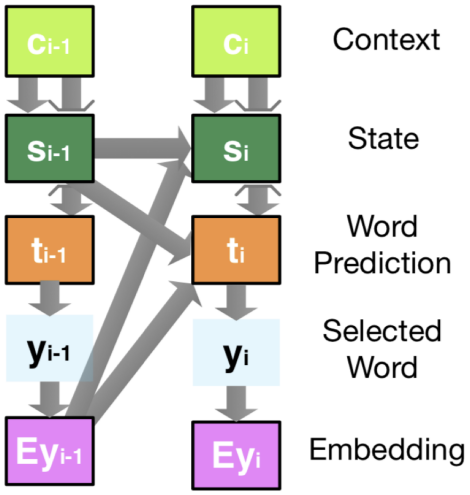
   SAN = ti, DOO = ce, MIO = bu, DET = ma, FAN = ho

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 $s_i = f(s_{i-1}, Ey_{i-1})$?



(a) Missing the previous state s_{i-1} will make this not recurrent. There will be no memory keeping track of previous state, though one might be able to depend on y_{i-1} for that. (b) There is no explicit information about the previously translated word, though this info may be captured by s_{i-1} somewhat. (c) The decoder is completely disconnected from the encoder, so it would just hallucinate without conditioning on the source side. Comment: generally any explanation that seems plausible is fine; the goal is to check whether students understand what is the motivation of each connection in the neural net.

Extra Space