First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2025
Johns Hopkins University
Co-ordinator: Philipp Koehn

2 October 2025

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Morphology *10 points*

1. Give an example an English word where the prefix has undergone a sound-shift from its canonical form: *(2 points)* e.g. impossible, illogical
2. Is Turkish a POLYSYNTHETIC or AGGLUTINATIVE language (circle one). *(2 points)* e.g. agglutinative
3. Is it easier to translate from a highly inflected language to a minimally inflected language or vice-versa? Explain in detail why. *(6 points)* ..

Q2. Syntax *20 points*

Formal Language Theory

1. We defined a formal language $\mathcal{L}$ as a set of strings, where each string is composed of zero or more symbols, each of which is drawn from an alphabet, Σ.

   a. What is the notation for *all* possible strings drawn from Σ? *(2 points)*

   b. What does ϵ denote? *(2 points)*

   c. Define alphabets that could be used for the following languages:

$$\begin{aligned} \mathcal{L}_1 &= \{\text{the set of English words}\} \\ \mathcal{L}_2 &= \{\text{the set of words in Russian}\} \\ \mathcal{L}_3 &= \{\text{all words in all human languages}\} \end{aligned}$$

   Σ_1 =

   Σ_2 =

   Σ_3 = *(6 points)*

All strings is Σ^*w or "sigma star". ϵ denotes the empty string. The Σs for the three languages are the set of Latin-1 characters (ASCII isn't enough, since English words can have diacritics), the set of cyrillic characters, and the entire unicode character set.

2. We spent some time in class discussing the move from representing languages as explicit sets, to representing languages as the set of strings produced by some generative process, defined by lists of rules. The Chomsky hierarchy defines four language classes of increasing power, corresponding to four constraints on the rule format. In the box below, A, B, and C are nonterminal symbols, while α, β, and γ represent a mix of nonterminal and terminal symbols (with a length of at least one). **Note**: one entry in each column is a fake.

   a. Cross out the fake item in each column. *(1 point)*

   b. Draw a line connecting each remaining item in the *rule format* column with the corresponding entry in the *language class* column. *(4 points)*

   c. For the two least powerful language classes (discussed in class), provide an informal example or description of a language that can be captured with rules from that class. *(4 points)*

   d. Put a star next to the most powerful language class, and describe informally what its rules mean. *(1 point)*

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

(a) $\alpha \rightarrow \beta$ and "context-friend" are fake. (b) In order, the rule formats are regular, context-free, context-sensitive, fake, and recursively enumerable. (c) The weakest class is the class of regular languages. One regular language is the set of all integers, which can easily be captured with a rule that generates a digit and another rule that generates both a digit and the nonterminal. The next-weakest class is the class of context-free languages. Note that all regular languages are also context-free, so any regular language is also an example context-free language. The classic example of a

language that is context-free but *not* regular is the set of palindromes. (d) The most power language class is the class of recursively enumerable languages. The corresponding rule ($\alpha A \beta \rightarrow \gamma$ means that any sequence of terminals and nonterminals can be rewritten as anything; i.e., this situation is the absence of any constraint on the rules.

3. 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.

Q3. Semantics *15 points*

	W_1	W_2	W_3	W_4	W_5	W_6
D_1	6	0	9	0	0	3
D_2	7	0	0	8	0	2
D_3	0	8	0	0	11	3
D_4	0	0	0	10	9	2

1. Consider the above matrix of co-occurrence counts (documents by words). After the application of an appropriate distributional-semantic method (a matrix factorization like SVD, a Bayesian model like LDA, etc), which pair of words, and which pair of documents, will show dramatic increase in similarity? *(10 points)* W_1 and W_5, initially with zero similarity, will move closer due to their association with W_4, and that shift will in turn draw D_2 and D_4 closer together.

2. Would it be more plausible for W_1 to be *monosemous* or *polysemous*, and why? *(5 points)* Polysemous, because while it shifts in sense closer to W_5 via D_2. its sense in D_1 appears distinct in terms of contextual co-occurrence.

Q4. Deep Learning *20 points*

1. What was one of the biggest problems with Recurrent Neural Networks? *(5 points)* Use your best judgment here. What I told them in class (multiple times) was vanishing gradients. They will probably have some pretty good other answers that will be worth points as well.

2. Language ID, introduced in the Deep Learning Lecture, is what type of Machine Learning Problem? *(10 points)* I'm looking for "classification" and "supervised learning". I imagine this will get a range of answers beyond these two. So either of the 2 are correct, plus possibly many more

3. What changed about a decade ago to usher in the era of machine learning? Please only give ONE answer. There are multiple and any ONE that is correct will get full points. *(5 points)* I'm looking for either "size of data" or "compute/gpus". If they give another answer that makes sense, sure. If they give more than one answer, give a 0

Q5. Information Retrieval *10 points*

1. Why are cross-encoders usually only used as rerankers? *(2 points)* Because the documents cannot be preprocessed or indexed like bi-encoders do.

2. What is the common approach for creating a numerical value for pointwise ranking using a generative model? *(3 points)* Take the softmax between the probability of generating yes over no (or any other reasonable pairs). Other acceptable answers could be (a) training a head/pooler on the (last) hidden states to produce a numerical score, or (b) prompting the model to give a graded rating. However, since option (b) is likely to produce many ties among the candidate documents, this answer can only receive 2 points.

3. What are the essential ingredients for using Cranfield evaluation? *(3 points)* A fixed document collection (1 point), a set of topics/queries (1 point), and relevance judgments/labels for a subset of documents for each topic (1 point).

4. Why is retrieval still important for LLM assistant systems? *(2 points)* Because the generation model has limitations in knowledge, particularly events after the model was trained (1 point), and retrieval provides grounded information for improving the trustworthiness of the system output (1 point).

Q6. Large Language Models *10 points*

You want to build a sentiment detection model (i.e., determining if a text's sentiment is either positive or negative) using either a BERT-style model or a Large Languege Model (LLM).

1. What data do you need to set this up as a supervised learning task? *(2 points)* Text with annotation as positive/negative

2. How would you build such a model based on a pre-existing BERT-style model? Provide some details about the model architecture of your added components.
   (4 points) Encode the text with the BERT model, and then add a custom classifier, such as a max-pooling layer, followed by a number of feed-forward layers.

3. How would you use a Large Language Model to handle this task? Name at least three methods that you may deploy. *(4 points)* In-context learning: describe the task to the LLM. Few-shot learning: provide examples in prompt. Fine-tune the model.

Q7. Preference Training *15 points*

You want to build a short story generation bot with the help of a Large Language Model. You want the stories to be engaging, creative, and have a twist. Your plan is to use reinforcement learning with human feedback.

1. You have access to human annotators. What will you ask them to do?*(4 points)* Rate stories generated by the model using the three criteria.
2. What model will you train with the human-annotated data? *(3 points)* Reward/Preference model.
3. You will be using Direct Preference Optimization. The formula for this is:

$$\text{Loss}(x, y_w, y_l) = \log \text{sigmoid}\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)$$

Explain in a sentence each, what the terms in this equation refer to, given your setup. *(8 points)*

(a) y_w: Story that the reward model prefers

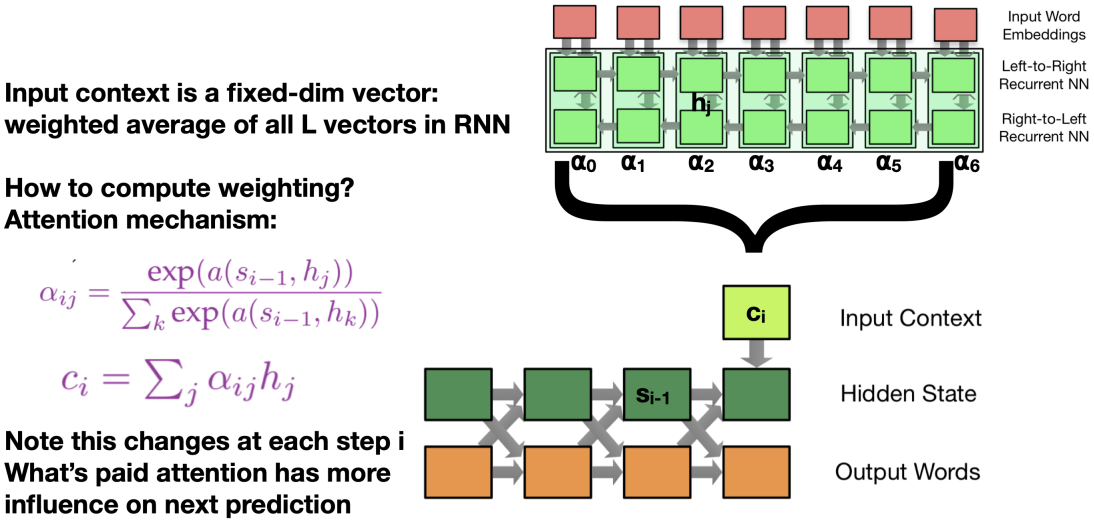
(b) y_l: Story that the reward model disprefers

(c) π_θ: The adapted LLM you are training

(d) π_{ref}: The original LLM

Q8. Machine Translation *10 points*

Answer the following questions with one or two sentences each, using the following figure of a RNN with Attention architecture if needed.



1. What would happen to the Machine Translation (MT) system output if the input context c_i is removed or set to zero? *(4 points)* The MT system will not be using any input information from the encoder. It might just hallucinate and not translate.

2. Explain what is the intuition behind α_{ij} in the equation $c_i = \sum_j \alpha_{ij} h_j$? *(4 points)* α_{ij} represents the relevance/importance of input word at position j for the decoding step i. Words with high α_{ij} will be emphasized in weighted average c_i and therefore will impact the decoding of the next word.

3. Suppose we replaced all h_j (contextual embedding) in the α_{ij} and c_i equations with the input word embeddings instead. What do you think will happen? *(2 points)* All reasonable answers will be accepted. For example: The MT system might still work since we are still giving it word embeddings, but it might perform worse since the embeddings are static. Alternative example: The MT system will fail because it can no longer contextual embeddings are extremely important. Yet another example: Transformers do not use RNN encoders, so maybe it will be ok given sufficient training.

Extra Space