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)*

2. Is Turkish a POLYSYNTHETIC or AGGLUTINATIVE language (circle one). *(2 points)*

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)*

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	

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

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)*

2. Would it be more plausible for W_1 to be *monosemous* or *polysemous*, and why? *(5 points)*

Q4. Deep Learning *20 points*

1. What was one of the biggest problems with Recurrent Neural Networks? *(5 points)*

2. Language ID, introduced in the Deep Learning Lecture, is what type of Machine Learning Problem? *(10 points)*

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)*

Q5. Information Retrieval *10 points*

1. Why are cross-encoders usually only used as rerankers? *(2 points)*

2. What is the common approach for creating a numerical value for pointwise ranking using a generative model? *(3 points)*

3. What are the essential ingredients for using Cranfield evaluation? *(3 points)*

4. Why is retrieval still important for LLM assistant systems? *(2 points)*

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)*

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)*

3. How would you use a Large Language Model to handle this task? Name at least three methods that you may deploy. *(4 points)*

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)*

2. What model will you train with the human-annotated data? *(3 points)*

3. You will be using Direct Preference Optimization. The formula for this is:

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

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

   (a) y_w:

   (b) y_l:

   (c) π_θ:

   (d) π_{ref}:

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.

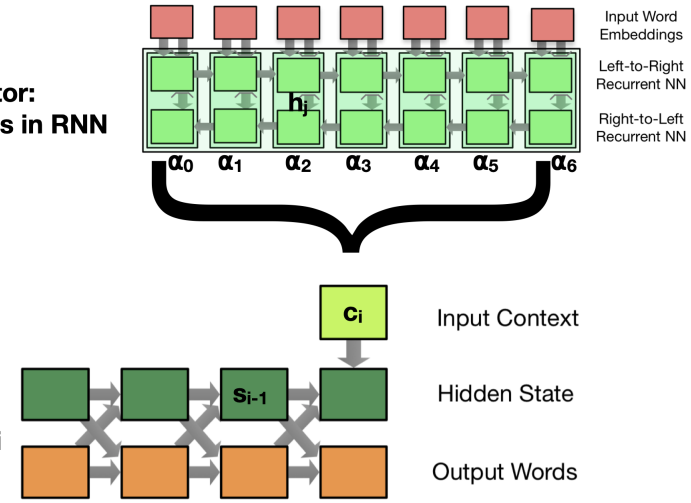
Input context is a fixed-dim vector: weighted average of all L vectors in RNN

How to compute weighting? Attention mechanism:

$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_k \exp(a(s_{i-1}, h_k))}$$

$$c_i = \sum_j \alpha_{ij} h_j$$

Note this changes at each step i What's paid attention has more influence on next prediction



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)*

2. Explain what is the intuition behind α_{ij} in the equation $c_i = \sum_j \alpha_{ij} h_j$? *(4 points)*

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)*

Extra Space