+1/1/60+

EPFL

1

Prof. Antoine Bosselut
Modern Natural Language Processing – CS-552
09.04.2025 from 11h30 to 13h00
Duration : 90 minutes

Midterm Practice Set

SCIPER : **111111**

Do not turn the page before the start of the exam. This document is double-sided, has 7 pages, the last ones are possibly blank. Do not unstaple.

- This is a closed book exam. Non-programmable calculators are allowed. No other electronic devices of any kind are allowed.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet if you have one; place all other personal items below your desk.
- You each have a different exam.
- This exam has **multiple-choice** questions of varying difficulty. Each question is worth **one point**.
- Each question has **exactly one** correct answer. For each question, mark the box corresponding to the correct answer. You are not expected to get every question right even for the best grade.
- Only answers in this booklet count. No extra loose answer sheets. You can use the blank pages at the end as scrap paper.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question turns out to be wrong or ambiguous, we may decide to nullify it.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
☒ ☑ ■	☐	☐ ▭

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

▨ ☒ ▢ · ▨ ▢

Question 1 Consider the equations for a standard LSTM cell:

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\
f_t &= \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\
o_t &= \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\
\tilde{c}_t &= \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \phi(c_t)
\end{aligned}
$$

In the equations above, which term explicitly represents the memory component that enables the LSTM to retain **long-term information** across timesteps?

- ☐ Output gate o_t
- ☐ Hidden state h_t
- ☐ Cell state c_t
- ☐ Candidate cell state $\tilde{c}_t$
- ☐ Input gate i_t

Question 2 BERT introduces a special token, [CLS], at the beginning of every input sequence. Which of the following statements best describes the purpose of the [CLS] token?

- ☐ It serves as a placeholder whose final hidden representation acts as a holistic sequence-level embedding, typically used for classification or next-sentence prediction tasks.
- ☐ It serves primarily to separate multiple sentences within the same input (the same role as [SEP] does).
- ☐ It simply marks sentence boundaries and carries no trainable embeddings of its own.
- ☐ It marks the exact midpoint of the input sequence to ensure balanced bidirectional attention.
- ☐ It is used only during masked language modeling and is dropped for downstream tasks.

Question 3 From the following set of models: {`ELMo`, `BERT`, `GPT`, `BART`, `T5`}, which group can each be directly used for both classification and generation tasks (without any modifications)?

- ☐ ELMo, BERT
- ☐ BERT, GPT
- ☐ BART, T5
- ☐ BERT, GPT, T5
- ☐ ELMo, BART, GPT

Question 4 Which of the following best defines semantics-encoding embeddings of words?

- ☐ A learned transformation of one-hot vectors into fixed-length random projections that improves computational efficiency.
- ☐ A technique that clusters words based on their dictionary definitions, ensuring that words with similar meanings always have identical representations.
- ☐ A representation of words as vectors, where the relative distance encodes semantic similarity based on co-occurrence patterns.
- ☐ A representation of words as sparse vectors in a high-dimensional space, ensuring that each word has a unique but unrelated position in the space.

Question 5 Which of the following is **FALSE** regarding autoregressive natural language generation?

- ☐ At each step during inference, the model predicts a probability distribution over the vocabulary space.
- ☐ The next generated token is the one with maximum probability as predicted by the model.
- ☐ We train a model to maximize the likelihood of the next token given the preceding tokens.
- ☐ Beam Search is more likely to generate a more probable sequence than greedy argmax decoding.

Question 6 Which of the following best describes **Chain-of-Thought (CoT) prompting** in large language models?

- ☐ A reinforcement learning technique that optimizes a language model's response generation using reward signals based on coherence and logical correctness.
- ☐ A prompting technique where multiple examples are given in the prompt to guide the model towards correct predictions through imitation learning.
- ☐ A method where the model is fine-tuned on logical reasoning tasks to improve its structured decision-making capabilities.
- ☐ A reasoning-based prompting method that encourages the model to break down complex problems into intermediate reasoning steps before producing a final answer.

Question 7 Which of the following claims are **NOT TRUE** about the *perplexity* metric?

(a) Easy to implement.

(b) Using base-2, base-3 and base-e will lead to the same perplexity score.

(c) Can be cheated by predicting low-frequency tokens.

(d) Can be very sensitive to high frequency tokens.

(e) The perplexity score can be 0.

- ☐ A and B.
- ☐ B and E.
- ☐ C and E.
- ☐ A and E.
- ☐ B, D and E
- ☐ B and C.
- ☐ C, D, E
- ☐ B, C and D.

Question 8 Applying N-gram language model, Fixed-window language model and a RNN language model on the same training dataset, which one will have the largest model size (i.e. greatest number of parameters)?

- ☐ Cannot be determined from the given information
- ☐ Fixed-window language model
- ☐ RNN
- ☐ N-gram language model

Backpropagation

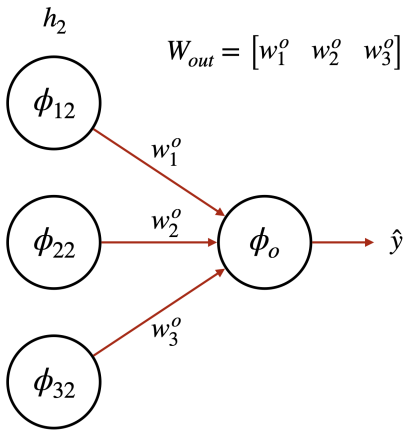

Figure 1: FFN Backpropagation

Question 9 According to Figure 1, the forward pass includes three steps:

$$u = W_{out}h_2 = w_1^o \times \phi_{12}(\cdot) + w_2^o \times \phi_{22}(\cdot) + w_3^o \times \phi_{32}(\cdot)$$
$$\hat{y} = \phi_0(u)$$
$$L(\hat{y}, y) = -y \ln \hat{y}$$

Given that ϕ_0 is the `ReLU` function, $W_{out} = [0.53 \quad -0.21 \quad 1.04]$, $h_2 = [0.36 \quad 1.02 \quad 5.18]^T$, $y = 4.20$, what is the partial gradient $\frac{\partial L(\hat{y},y)}{\partial \phi_{32}(\cdot)}$?

- [ ] 4.06
- [ ] 0.81
- [ ] 0.41
- [ ] -4.06
- [ ] -0.81
- [ ] -0.41

Question 10 Now given that the loss function is the L2 loss $L(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2$, what is the partial gradient $\frac{\partial L(\hat{y},y)}{\partial w_3^o}$?

- [ ] 1.21
- [ ] 0.42
- [ ] 6.03
- [ ] -1.21
- [ ] -0.42
- [ ] -6.03