ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (CS–431)
Bosselut, A., Chappelier, J.-C. & Rajman, M.

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (CS–431)

Fall 2024 — Solution of the exam

Friday, January 24th, 2025.

BLANK PAGE. DO NOT WRITE HERE.

QUESTION I : Back to basics [45 pts]

① **(a) [4 pts]** How does "*medical*" compare to "*camelid*" using a bigram model of characters and using a trigram model?

In both cases, provide your answers as formulas using only model parameters and with the fewer possible terms.

(b) [2 pts] Is one of these two models better than another on this comparison? Justify your answer.

For bigrams (in blue, the common terms):

$$P(\text{medical}) = P(\text{me})\,\frac{P(\text{ed})}{P(\text{e})}\,\frac{P(\text{di})}{P(\text{d})}\,\frac{P(\text{ic})}{P(\text{i})}\,\frac{P(\text{ca})}{P(\text{c})}\,\frac{P(\text{al})}{P(\text{a})} = K \cdot \frac{P(\text{ed})\,P(\text{di})\,P(\text{ic})\,P(\text{al})}{P(\text{d})\,P(\text{c})}$$

$$P(\text{camelid}) = P(\text{ca})\,\frac{P(\text{am})}{P(\text{a})}\,\frac{P(\text{me})}{P(\text{m})}\,\frac{P(\text{el})}{P(\text{e})}\,\frac{P(\text{li})}{P(\text{l})}\,\frac{P(\text{id})}{P(\text{i})} = K \cdot \frac{P(\text{am})\,P(\text{el})\,P(\text{li})\,P(\text{id})}{P(\text{m})\,P(\text{l})}$$

where $P(x) = \sum_y P(xy)$, with x and y single chars.

For trigrams (these are not the same $P()$ as above):

$$P(\text{medical}) = P(\text{med})\,\frac{P(\text{edi})}{P(\text{ed})}\,\frac{P(\text{dic})}{P(\text{di})}\,\frac{P(\text{ica})}{P(\text{ic})}\,\frac{P(\text{cal})}{P(\text{ca})}$$

$$P(\text{camelid}) = P(\text{cam})\,\frac{P(\text{ame})}{P(\text{am})}\,\frac{P(\text{mel})}{P(\text{me})}\,\frac{P(\text{eli})}{P(\text{el})}\,\frac{P(\text{lid})}{P(\text{li})}$$

where $P(xy) = \sum_z P(xyz)$, with x, y and z single chars.

Trigrams are more discriminating in this case since these two words have no trigram in common but have two bigrams in common ("*me*" and "*ca*").
However, trigrams have more parameters, thus harder to estimate and more sensitive to sparse data. If there are enough training data, trigrams will be better, otherwise bigrams might be a better option.

Comment: Several students didn't go as far as using the parameters.

② **[4 pts]** On a corpus of 3'147'235 occurrences of 237 unique characters, how would a parameter of a trigram character model be estimated using a Dirichlet prior with a uniform parameter set to $5 \cdot 10^{-4}$ if its Maximum-Likelihood estimate on the same corpus is $\widehat{p}$?

Justify your answer.

A corpus of 3'147'235 characters contains 3'147'233 trigrams.

The MLE is then $\widehat{p} = \frac{N}{3'147'233}$, where N is the number of occurrences in the corpus of the considered trigram.

The total number of possible trigrams with this alphabet is 237^3.

The additive smoothing via Dirichlet prior on the same trigram is then $\dfrac{N + 5 \cdot 10^{-4}}{3'147'233 + 5 \cdot 10^{-4} \times 237^3}$,
i.e. $\dfrac{3'147'233\,\widehat{p} + 5 \cdot 10^{-4}}{3'147'233 + 5 \cdot 10^{-4} \times 237^3}$

③ **[15 pts]** Consider the following sentence:

they mournfully bark at the fair light moon

and an order-1 HMM for Part-of-Speech tagging with the following parameters (not exhaustive, but no missing information to solve the question):

they: Pron: $2.1 \cdot 10^{-5}$
mournfully: Adv: $3.2 \cdot 10^{-5}$
bark: N: $2.5 \cdot 10^{-5}$, V: $1.4 \cdot 10^{-5}$
at: Prep: $3.7 \cdot 10^{-5}$
the: Det: $4.5 \cdot 10^{-5}$
fair: Adj: $3 \cdot 10^{-5}$, Adv: 10^{-5}, N: $2 \cdot 10^{-5}$
light: Adj: $5 \cdot 10^{-5}$, N: $3 \cdot 10^{-5}$, V: $7 \cdot 10^{-5}$
moon: N: $1.6 \cdot 10^{-5}$

	Adj	Adv	Det	N	Prep	Pron	V
Adj	0.7	0.02	0.01	**x**	0.04	0.03	0.1
Adv	0.002	0.08	0.09	0.001	0.07	0.06	0.6
Det	0.4	0.1	**0**	0.5	**0**	**0**	**0**
N	0.2	0.05	0.02	**y**	0.15	0.08	0.3
Prep	0.1	0.06	0.4	0.18	**0**	0.16	0.1
Pron	0.1	0.3	**0**	0.2	**0**	**0**	0.4
V	0.14	0.16	0.13	**z**	0.15	0.11	0.12

(a) [10 pts] Assume that $5y > 6x$. Provide the tightest possible condition(s) between x, y and z so that the tag of "*light*" in the most probable sequence of tags for the above sentence is N.

(b) [5 pts] If these conditions are fulfilled, what is the most probable sequence of tags for the above sentence?

Fully justify your answers (most of the points are for the justifications.)

First notice that the above table contains transition probabilities from row tag to column tag (look at, for instance, the sum of the first column).

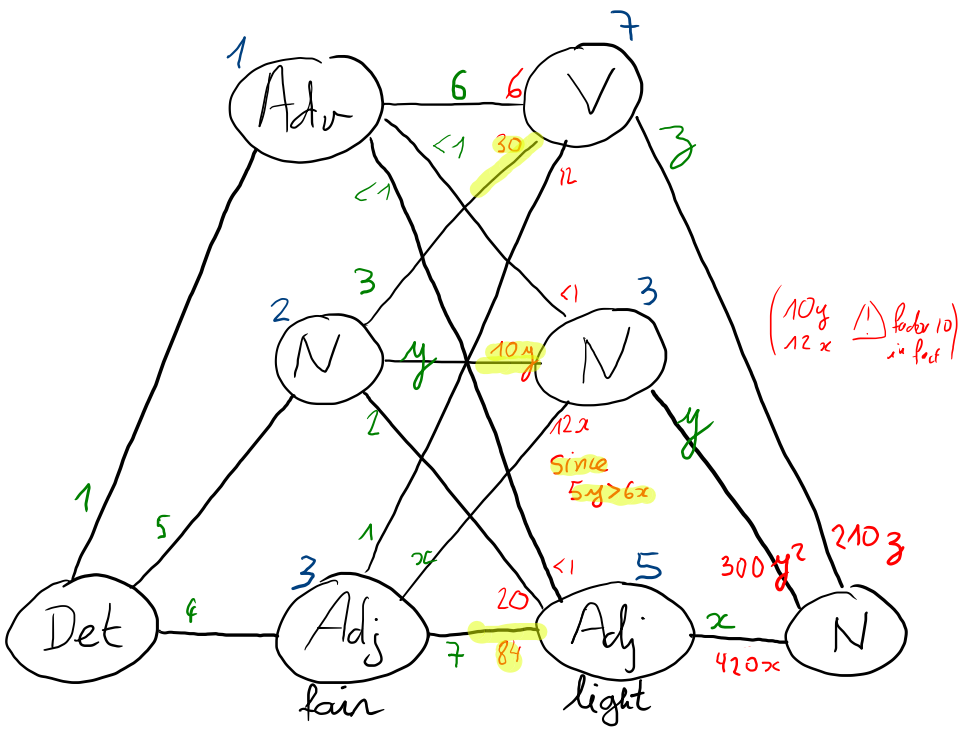
Second, we have quite some non-ambiguous words:

they	mournfully	bark	at	the	fair	light	moon
Pron	Adv		Prep	Det			N

which makes the optimization of the two ambiguous parts *independent*.

Maximizing the first part ("*bark*") is not difficult and can be done in your head without any advanced algorithm:
$P(\texttt{N}|\texttt{Adv}) \cdot P(\texttt{Prep}|\texttt{N}) \cdot P(\text{bark}|\texttt{N})$
compared to
$P(\texttt{V}|\texttt{Adv}) \cdot P(\texttt{Prep}|\texttt{V}) \cdot P(\text{bark}|\texttt{V})$;
i.e. $0.001 \times 0.15 \times 2.5$ compared to $0.6 \times 0.15 \times 1.4$, leading to V.

For the second part, use the Viterbi algorithm (rather than brute force):



(also assumed: $y > P(\text{N}|\text{Adv}) = 0.001$, which makes sense syntactically).

The conditions for "*light*" to be tagged as `N` are thus:

$$3y^2 > 4.2x \text{ (i.e. } y^2 > 1.4x) \quad \text{and} \quad 3y^2 > 2.1z \text{ (i.e. } y^2 > 0.7z)$$

$$(\text{and } x \leq 0.1,\ y \leq 0.2,\ z \leq 0.19)$$

The most probable sequence is then:

they	mournfully	bark	at	the	fair	light	moon
Pron	Adv	V	Prep	Det	N	N	N

④ **[4 pts]** Assume we want to classify sentences according to two classes, `Poetry` and `Technical`. What is the probability for the following sentence:

they mournfully bark at the fair light moon

to be classified as each of the two above classes, using a (1-gram) Naive Bayes system, trained on a corpus where:

- "*they*", "*at*" and "*the*" are removed by preprocessing;
- 40% of the sentences are in `Poetry` (the rest being in `Technical`, no sentence in common);
- the relative frequency of each of the following words for each of the two classes are:

	`Poetry`	`Technical`
mournfully	10^{-5}	$0.375 \cdot 10^{-5}$
bark	$3 \cdot 10^{-5}$	$6.75 \cdot 10^{-5}$
fair	$4 \cdot 10^{-5}$	$6 \cdot 10^{-5}$
light	$9 \cdot 10^{-5}$	$31.5 \cdot 10^{-5}$
moon	$8 \cdot 10^{-5}$	$8 \cdot 10^{-5}$

- the relative frequency of each of the two classes for each of the following words are:

	`Poetry`	`Technical`
mournfully	0.8	0.2
bark	0.4	0.6
fair	0.5	0.5
light	0.3	0.7
moon	0.6	0.4

Provide your answer as a product of numbers (no need for a final value) and *fully justify* your answer (explaining assumptions where needed).

For class `C`:

$$P(\texttt{C}|\texttt{words}) \propto P(\texttt{C}) \times \prod_{\texttt{words}} P(\texttt{word}|\texttt{C})$$

Comment: Too many write an equalitiy here, rather than "proportional to".

For class `Poetry`: $\propto 0.4 \times 1 \times 3 \times 4 \times 9 \times 8$

For class `Technical`: $\propto 0.6 \times 0.375 \times 6.75 \times 6 \times 31.5 \times 8$

(with the same proportional coefficient)

Comment: There are few students who did use the $P(C|\text{word})$ values rather than $P(\text{word}|C)$ (on what do the values sum up to 1?).

⑤ **[13 pts]** We now want to build a new classification method for `Poetry` against `Technical`, that combines Naive Bayes classification and Part-of-Speech tagging using HMMs.

(a) [1 pt] What is the main assumption made for Naive Bayes classification?
Express it as a mathematical formula and explain how it is useful for the considered task.

(b) [2 pts] What are the two main assumptions made for applying HMMs to Part-of-Speech tagging?
Express them as a mathematical formulas and explain how they are useful for the considered task.

(c) [3 pts] Which of the two assumptions proposed in **(b)** do you propose to modify in the same spirit as in **(a)** so as to build a classifier that combines Naive Bayes classification and Part-of-Speech tagging using HMMs?
Express your proposed assumption as a mathematical formula.

(d) [2 pts] Using the assumption you proposed in **(c)**, how is the classification problem expressed mathematically?

(e) [5 pts] Do you think such a model could provide better classification results than the standard Naive Bayes model?
Fully justify your answer and provide concrete examples to support your claims.

(a) Conditional independence of words(/features) knowing class:

$$P(w_1^n|C) = \prod_i P(w_i|C)$$

This is useful to simplify the whole formula using Bayes rule:

$$P(C|w_1^n) \propto P(C) \cdot P(w_1^n|C) = P(C) \prod_i P(w_i|C)$$

There are thus much less parameters in the final model than in the original formulation.

(b) Both assumptions are *limited lexical conditioning* (leading to emissions probabilities) and *limited syntactic dependencies* (Markov assumption on tag sequences):

$$P(w_i|w_1, ..., w_{i-1}, t_1, ..., t_n) = P(w_i|t_i)$$

$$P(t_i|t_1, ..., t_{i-1}) = P(t_i|t_{i-k}, ..., , t_{i-1})$$

Similarly to Naive Bayes assumption, these are useful to simplify the whole formula introducing way less parameters than in the original formulation ($P(w_1^n, t_1^n)$; this is a generative model).

(c) The easiest seems to be the *limited lexical conditioning*, simply introducing the classes:

$$P(w_i|C, w_1, ..., w_{i-1}, t_1, ..., t_n) = P(w_i|t_i, C)$$

From the Naive Bayes point of view, this simply adds the tag in the conditioning: $P(w_i|t_i, C)$ instead of $P(w_i|C)$.

(d) The classification problem expressed in probabilistic terms aims at maximizing (over possible values for C) $P(C|w_1^n)$, which using Bayes Rules turns into maximizing $P(w_1^n, C)$, which using our new limited lexical conditioning assumptions and keeping the limited syntactic dependencies unchanged leads to (for an order-1 HMM):

$$P(w_1^n, C) = P(C) \sum_{t_1,...,t_n} P(t_1)P(w_1|t_1, C) \prod_{i=2}^{n} P(t_i|t_{i-1})P(w_i|t_i, C)$$

Actually, keeping the limited syntactic dependencies unchanged, makes in fact another probabilistic assumption, which is that tag sequences are independent from the class. If we really want to fully generalize, then we have to keep the class conditioning in the tag sequence, which for `Poetry` against "usual text" makes even more sense:

$$P(w_1^n, C) = P(C) \sum_{t_1,...,t_n} P(t_1|C)P(w_1|t_1, C) \prod_{i=2}^{n} P(t_i|t_{i-1}, C)P(w_i|t_i, C)$$

(e) This is difficult to say without more context:

- on one hand, and especially for poetry where both the vocabulary and the syntactic structure are different from usual text, it makes sense to condition the words (and tag sequence) to the class; this makes the model more able to capture these differences;
- on the other hand, this introduces twice as many parameters (which are already many); to benefit from such a richer model, we should ensure we have enough data to train it properly.

⑥ **[3 pts]** Here, we consider the evaluation of a document retrieval system for 4 queries and 6 documents.

Document 1 is relevant for queries 1 and 2;
document 2 is relevant for query 3 only;
document 3 is relevant for queries 2 and 4;
document 4 is relevant for query 4 only;
document 5 is relevant for queries 1, 3 and 4;
document 6 is relevant for queries 2 and 4.

For query 1, the system classifies the documents in the following order (best document first):
5 1 | 3 6 4 2; (2 relevant documents)
for query 2: **6** 4 | **1** 2 **3** 5; (3 relevant documents)
for query 3: 4 **2** | **5** 1 6 3; (2 relevant documents) and
for query 4: 1 2 | **6 5 4 3**. (4 relevant documents)

Compute P@2 for each query and R-precision.

Justify your answers.

a) Before the desired rank (2 here; marked above) the order does not matter, only the number of relevant document (above colored in bold red) before that rank; so we get:
for query 1: $2/2 = 1$
for query 2: $1/2 = 0.5$
for query 3: $1/2 = 0.5$; and
for query 4: $0/2 = 0$.

b) We average: P@2 for q1, P@3 for q2, P@2 for q3 and P@4 for q4:

$$\text{R-Prec} = \frac{1}{4}\left(1 + \frac{2}{3} + \frac{1}{2} + \frac{1}{2}\right) = \frac{16}{24} = \frac{2}{3}$$

QUESTION II : A concrete job [40 pts]

You are a physician in CHUV's diagnostic department, dedicated to identifying the diseases behind your patients' symptoms. Over the years, as the number of patients has grown exponentially, your department has struggled to keep up with the increasing demand. This has led you to wonder: how can I develop an automated medical assistant to help the team provide faster and more accurate diagnoses?

This medical assistant will receive the clinical note $\mathbf{X}$ of a patient as input and produce an initial diagnosis $\mathbf{Y}$ (in long-form text with an explanation). You decide to use a sequence-to-sequence model for this task, and to use transformers as both your encoder and decoder.

① **[2 pts]** You quickly realize that the self-attention component of the transformer brings a crucial downside compared to other models such as RNNs. Briefly describe this downside and propose a solution to overcome it.

- Self-attention provides no word order information.
- Position Embedding: Add an additional embedding to the input word that represents a position in the sequence

② **[1 pt]** You decide to train your model on a dataset of clinical notes, paired with doctor-written diagnostic reports. For any clinical note $\mathbf{X}$, and diagnostic report $\mathbf{Y} = [y_0,\ y_1,\ \cdots,\ y_t,\ \cdots,\ y_T]$, state the objective you would **minimize** to train your language model.

Minimize

$$-\log P(y_0, ..., y_T|X) = -\log P(y_0|X) - \sum_{t=1}^{T} \log P(y_t|y_0, ..., y_{t-1}, X)$$

Consider only the decoder in your sequence-to-sequence model. Let's take a look at how the transformer model will process a sequence. For the following questions, assume you have a short sequence with four tokens with the following input embeddings:

$$\text{blood} = \left[\ 2,\ 1\ \right] \qquad \text{pressure} = \left[\ 1,\ 1\ \right] \qquad \text{is} = \left[\ 0,\ 1\ \right] \qquad \text{good} = \left[\ 2,\ 0\ \right]$$

And your model has the following position embeddings:

$$p_1 = \left[\ -1,\ -1\ \right] \qquad p_2 = \left[\ -1,\ 0\ \right] \qquad p_3 = \left[\ 1,\ 0\ \right] \qquad p_4 = \left[\ 0,\ 1\ \right]$$

③ **[8 pts]** Using scaled dot product attention, what is the attention distribution over key vectors for the token "*good*" as the query in the first attention layer?
Assume that W^K and W^V are identity matrices and

$$W^Q = \begin{pmatrix} \sqrt{2}\ \ln(4) & 0 \\ 0 & \sqrt{2}\ \ln(4) \end{pmatrix}$$

Justify your answer and provide all the steps of your computation.

1. Integrate Position Embedding:

$$\text{blood} = \left[\ 2,\ 1\ \right] + \left[\ -1,\ -1\ \right] = \left[\ 1,\ 0\ \right]$$
$$\text{pressure} = \left[\ 1,\ 1\ \right] + \left[\ -1,\ 0\ \right] = \left[\ 0,\ 1\ \right]$$
$$\text{is} = \left[\ 0,\ 1\ \right] + \left[\ 1,\ 0\ \right] = \left[\ 1,\ 1\ \right]$$
$$\text{good} = \left[\ 2,\ 0\ \right] + \left[\ 0,\ 1\ \right] = \left[\ 2,\ 1\ \right]$$

2. Query Computation for "good":

$$Q(\text{good}) = W^Q \begin{bmatrix} 2, & 1 \end{bmatrix} = \sqrt{2}\ \ln(4) \begin{bmatrix} 2, & 1 \end{bmatrix}$$

3. Normalized Dot Products Computation

$$s(\text{blood}) = \frac{1}{\sqrt{2}} Q(\text{good}) \begin{bmatrix} 1, & 0 \end{bmatrix}^T = 2\ln(4)$$
$$s(\text{pressure}) = \frac{1}{\sqrt{2}} Q(\text{good}) \begin{bmatrix} 0, & 1 \end{bmatrix}^T = \ln(4)$$
$$s(\text{is}) = \frac{1}{\sqrt{2}} Q(\text{good}) \begin{bmatrix} 1, & 1 \end{bmatrix}^T = 3\ln(4)$$
$$s(\text{good}) = \frac{1}{\sqrt{2}} Q(\text{good}) \begin{bmatrix} 2, & 1 \end{bmatrix}^T = 5\ln(4)$$

4. Softmax Numerators:

$$\exp(s(\text{blood})) = 4^2 = 16 \qquad \exp(s(\text{pressure})) = 4^1 = 4$$
$$\exp(s(\text{is})) = 4^3 = 64 \qquad \exp(s(\text{good})) = 4^4 = 1024$$

5. Softmax Normalization:

$\text{sum} = 16 + 4 + 64 + 1024 = 1108$

Final Answers:

$$a(\text{blood}) = \frac{16}{1108} \qquad a(\text{pressure}) = \frac{4}{1108} \qquad a(\text{is}) = \frac{64}{1108} \qquad a(\text{good}) = \frac{1024}{1108}$$

Now that the transformer model can process the input text correctly, we can analyze how the model generates text.
On the provided extra loose-leaf sheet, we show the top-2 highest probability tokens in these distributions at each step (along with their **log probability**).

④ **[4 pts]** Compute the sequence loglikelihood for each possibly-generated sequence in the diagram and, on the provided extra loose-leaf sheet, *annotate* the boxes to the right of the diagram with these **loglikelihoods**.

-11.1, -13.1, -12.7, -9.8

-12.5, -7, -10.2, -8.1

-8.8, -11.1, -10.1, -9.1

-7.5, -5.7, -7.8, -10.4

⑤ For the following sub-questions, give your answer using the indices (1–16) in the diagram to the right of the figure.

(a) [1 pt] What is the optimal sequence? **Answer:** 14

(b) [1 pt] What is the sequence produced by argmax decoding? **Answer:** 15

(c) [3 pts] What is the **second-best** final sequence produced by beam search with a beam size of 2? **Justify** your answer by annotating the diagram (on the provided extra loose-leaf sheet).

Answer: 13

(d) [1 pt] Which sequence are you **least** likely to produce using top-k sampling with $k = 2$?

Answer: 2

(e) [3 pts] List the sequences that **cannot** be generated if you use top-p sampling with $p = 0.4$. **Justify** your answer.

Hints: assume $\ln(0.4) \simeq -0.91$; in top-p sampling, you sample from all tokens until the cumulative distribution **exceeds** the threshold.

Answer and justification: 5, 11, 13, 14

To provide your model the ability to access private patient records, you decided to enhance your model with Retrieval Augmented Generation (RAG). In simple words, you give the model more information by retrieving the most relevant documents using an embedding model. To represent documents in your retrieval-augmented system, you will rely on the sum of word embeddings as representations.

⑥ **[1 pt]** You currently have two choices to build your embedding model: CBOW or Skip-Gram. What's the key conceptual difference between these two methods?

CBOW learns to predict a missing word from the surrounding window of words.

Skip-Gram learns to predict the surrounding window of words from a given word.

⑦ **[6 pts]** You decide to use the Skip-Gram algorithm to train your word embeddings. You first test your training with a small vocabulary of five words and provide it the sequence of words "*the blood pressure is good*" with the following embeddings:

$$\text{the} = \begin{bmatrix} 0, & 1 \end{bmatrix} \quad \text{blood} = \begin{bmatrix} 0, & 0 \end{bmatrix} \quad \text{pressure} = \begin{bmatrix} 1, & 0 \end{bmatrix} \quad \text{is} = \begin{bmatrix} 1, & 1 \end{bmatrix} \quad \text{good} = \begin{bmatrix} 2, & 1 \end{bmatrix}$$

Compute the value of the loss function you need to minimize for the word "*pressure*", given a window size of 2. Leave the exponential terms in your computation.
Note: the output embeddings are the transpose of the input embeddings.

The window is: [the, blood, is good]

The input embedding is "presure": $\begin{bmatrix} 1, & 0 \end{bmatrix}$, the dot-product with which simply outputs the first component.

The projections (dot-products) of the five tokens are thus:

$$0 \quad 0 \quad 1 \quad 1 \quad 2$$

The loss function is:

$$\mathcal{L} = -\left(\ln P(\text{the}|\text{pressure}) + \ln P(\text{blood}|\text{pressure}) + \ln P(\text{is}|\text{pressure}) + \ln P(\text{good}|\text{pressure})\right)$$

where $P(w_x|w_t) = \text{softmax}(\mathbf{U}\, t)_x$, thus $\ln P(w_x|w_t) = x \cdot t - \ln S$ with $S = \sum_w \exp(w \cdot t)$.
In our case, $S = 1 + 1 + e + e + e^2$ and

$$\mathcal{L} = -\left(0 + 0 + 1 + 2 - 4 \ln S\right) = 4 \ln(2 + 2e + e^2) - 3$$

⑧ **[6 pts]** Before you finalize and submit your proposal to the hospital administrators, you need to validate if your trained medical copilot model can generalize to new cases. You first have to decide the evaluation metric you will use to grade the responses from the model. For the following three evaluation metrics, describe one advantage and one disadvantage:

(a) [2 pts] content overlap metrics:

Advantage: Fast and efficient, widely used, or cheap to compute

Disadvantage: Performs much worse for open-ended tasks. Does not measure more than word overlapping, such as semantics, which would penalize synonyms.

(b) [2 pts] model-based metrics:

Advantage: Can measure fine-grained semantics, more correlated with human judgment

Disadvantage: Difficult to interpret, requires training on labeled data

(c) [2 pts] human evaluation:

Advantage: Medical evaluation is highly nuanced, requiring a deep understanding across different scopes and specialties of medical knowledge. The evaluation is also highly context-dependent and requires ethical oversight, which is a strength of medical professionals with extensive training and clinical experience. Their expertise ensures that the recommendations provided by medical language models are accurate, fair, context-aware, and ethically sound.

Disadvantage: Expansive to recruit expert evaluators, time-consuming to design, execute, gather, and analyze; thus, overall, very difficult to scale. Expert evaluators can be subjective or biased, which can make the results inconsistent and incomparable across different studies.

⑨ **[3 pts]** You officially submitted your proposal for the medical copilot to the hospital administrators. The ethics committee of the hospital received the proposal and started a thorough review to look for any potential ethical issues. For each of the following ethical perspectives, briefly explain what kind of issues the committee might focus on during the review:

- harm from disinformation;
- leaking private patient information;
- bias and fairness.
- The ethics committee might focus on the potential for the model to produce inaccurate or misleading medical information, which could result in incorrect and inappropriate diagnoses, treatment recommendations, or drug prescriptions. Such misinformation could lead to significant harm to patients, including worsening health conditions or even death.

  The committee will likely examine the system's ability and accuracy to provide evidence-based recommendations, as well as the safeguards in place to ensure that medical professionals critically assess any information provided by the model.

  Additionally, they will consider the risks associated with relying on the model without appropriate oversight and the legal and professional repercussions for healthcare providers who rely on incorrect information generated by the model.
- The ethics committee might focus on the concern that the model could generate or leak private patient information, such as names, Social Security Numbers, addresses, or medical histories, especially if it is vulnerable to manipulation through methods like prompt injection, membership inference attacks, or LLM jailbreaking.

  The committee will review the privacy measures in place to prevent unauthorized access to sensitive information and evaluate compliance with regulations like HIPAA. They

will also look at how the system logs the human-model interactions, how it manages data retention and deletion, and what protocols are in place to handle any potential data breaches.

- The ethics committee might focus on the potential for the model to exhibit bias in its diagnoses or recommendations based on factors such as race, age, gender, sexuality, nationality, or socioeconomic status.

  For example, suppose the model has been trained on predominantly homogeneous datasets. In that case, it may produce biased outcomes that do not adequately account for variations in symptoms or treatment efficacy across different racial or ethnic groups.

  This could lead to disparities in the quality of care received by patients from underrepresented backgrounds. Such biases can result in unfair treatment, where certain patient groups either receive less effective medical advice or are systematically overlooked.

Quick example:
The model might incorrectly prioritize anxiety or non-cardiac causes for chest pain in women, reflecting learned biases rather than evidence-based medical practice. This can lead to unfair treatment outcomes, where women experience disparities in care quality, are more likely to experience diagnostic errors, and face a greater risk of adverse health outcomes.

QUESTION III : Science Fiction [35 pts]

Our exploration of exoplanets has finally yielded results: we have not always been alone in the universe. Evidence of an advanced civilization has been discovered on Proxima Centauri b (hereafter "PCb"), a planet orbiting the star Proxima Centauri approximately 4.24 light-years from Earth.

The primary characteristics of PCb are that it is a rocky planet exposed to intense radiation and stellar flares, yet it has liquid water on its surface. Its orbital period (its "year") is only 11.2 Earth days, and it is tidally locked: one side of the planet is always facing its star (permanent day), while the other side remains in perpetual darkness.

The discovered civilization appears to be extinct, but the probe sent to PCb retrieved a crystal containing 200 TB of optical storage consisting of sequences in what seems to be an alphabet of 325 symbols.

Building on recent advances in NLP, an international research team aims to investigate the hypothesis that the data collected represents Centaurian, the presumed language of the extinct civilization.

1. Linguistic Preliminaries

① **[2 pts]** What are the primary functions of a (human) natural language?
Restrict to the functions explicitly mentioned in the course.
For each mentioned function, provide an example of the kind of linguistic data that it could be associated with.

- **Communication** – Language serves as a tool for exchanging information between speakers.
  Example: A transcript of a conversation.
- **Representation of Knowledge** – Language encodes and preserves facts, ideas, and abstract concepts.
  Example: A physics textbook.

② **[5 pts]** How could one provide some evidence about the fact that the collected data could correspond to linguistic data comparable to the ones existing for human languages?
Justify your answer thoroughly. In particular, give some indication on how the requested evidence could be produced.

- **Evidence of Zipf's Law:** In human languages, word frequency follows a power-law distribution.
  Evidence Production: Compute the frequency of each symbol in the dataset and check if it follows Zipf's Law.
- **Multi-level Organization:** Human languages exhibit hierarchical structures, such as words forming phrases and sentences.
  Evidence Production: Identify patterns of symbol co-occurrence.
- **Distributional Semantics:** In human languages, word meaning can be inferred from context patterns.
  Evidence Production: Check whether certain sequences of symbols tend to appear in specific environments rather than being randomly distributed.

③ **[2 pts]** Human languages are known to be ambiguous and implicit. Briefly explain why this is the case.

Ambiguity allows expressive power.

Implicitness enables communication efficiency.

④ **[7 pts]** Evaluate whether ambiguity and implicitness could also apply to Centaurian. Regardless of your position on the matter, provide a detailed justification.

As implicitness (resp. ambiguity) increase conciseness (resp. expressive power), they likely exist in any advanced civilization's language, except in specific cases. A fully explicit language may be justified if explicitation errors pose great risks, while a non-ambiguous language would need alternative ways to handle partial understanding.

⑤ **[3 pts]** A human language is usually characterized by its syntax and semantics. Provide a brief definition of each of these two concepts, accompanied by at least one illustrative example. Make the distinction between semantics and pragmatics.

- **Syntax**: the rules that govern grammatical correctness, i.e. the correct **structure** and **arrangement** of words.
  Example: "*The cat sleeps*" is syntactically correct, whereas "*Sleeps Cat the*" is not.
- **Semantics:** the rules that define the **literal meanings** (i.e., independent of context).
  Example: "*Colorless green*" is semantically nonsensical.
- While Semantics deals with literal meaning, **Pragmatics** considers **contextual meaning**.
  Example: "*Paul saw Mary with a telescope*"; only pragmatics can decide who has a telescope.

⑥ **[4 pts]** Get inspiration from the linguistic capabilities exhibited by the LLMs to evaluate whether the concepts of syntax and semantics should necessarily apply to Centaurian.
Provide a detailed justification for your answer.

LLMs exhibit advanced **linguistic abilities** (e.g., generating well-formed sentences, capturing meaning) **without any built-in syntactic or semantic rules**.

Their performance suggests that **language use does not require *explicit* syntax or semantics**. Syntax and semantics, as formalized in linguistics, are **not necessarily inherent properties of a language** but rather **conceptual tools** that humans have developed to describe and analyze language. They might not necessarily apply to Centaurian.

2. A Bit of Research...

To advance their research, the international team decided to train a large language model (LLM) on a large corpus containing both texts in human languages and data from PCb.

⑦ **[4 pts]** Once the LLM is pre-trained, part of the international team decided to use it to translate from Centaurian. Indicate what should follow the pre-training for such an approach to be conceivable and whether you consider it as realistic or not.

After pre-training, the next step would involve fine-tuning on a parallel corpus. Since no such corpus exists, direct translation becomes highly speculative.

Alternative approaches might be considered: distributional methods might identify common statistical structures or cross-linguistic techniques might map Centaurian embeddings to human ones. However, their success would heavily depend on how closely Centaurian language (and world) resembles human ones.

⑧ **[4 pts]** Another part of the international team decided to explore a different approach: leveraging the conversational abilities of the pre-trained LLM to interact with a "Centaurian", typically through questions like "*Describe life on PCb*", "*What are Centaurians like?*", etc.
Does this approach seem substantially different from the previous one?
Justify your answer.

This approach is **fundamentally different** from the translation-based approach, as it does not require any **linguistic alignment** between Centaurian and human languages.

It aims to **generate plausible responses** based on the statistical properties of the training data.

It might produce **intriguing** insights.

However, there is no guarantee that they reflect **actual** Centaurian knowledge rather than **hallucinated constructs** shaped by the LLM's biases.

3. And to Conclude...

Ultimately, the international team compiled their findings into a set of 500 translations and 500 interactions. Each was assigned a binary score indicating whether a human expert found it plausible or not. With 51% of the cases deemed plausible, the obtained result has been published as such under the title: "*There Truly Was an Advanced Civilization on Proxima Centauri b!*"

⑨ **[4 pts]** What do you think of the published conclusion?
Justify your answer by critically analyzing the evaluation methodology used.
Focus on concepts that have been explicitly mentioned in the course and give priority to aspects that can be associated with some quantitative measure.

The conclusion is **not justified** due to a **flawed evaluation methodology**. The **lack of a Gold Standard**, the **absence of objective metrics**, and the **weak statistical significance** all undermine the credibility of the claim.