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)
Chappelier, J.-C., Rajman, M. & Bosselut, A.

Introduction to Natural Language Processing (CS–431)

Fall 2022 — Solution of the exam

Thursday, January 26th, 2023.

QUESTION I : The Daily Gazette [29 pt]

You have been publishing a daily column for the Gazette over the last few years and have recently reached a milestone — your 1000th column! Realizing you'd like to go skiing more often, you decide it might be easier to automate your job by training a story generation system on the columns you've already written. Then, whenever your editor pitches you a title for a column topic, you'll just be able to give the title to your story generation system, produce the text body of the column, and publish it to the website!

① **[2 pt]** You consider using either a transformer or a recurrent neural network (RNN) as the underlying model for your text generator. Assuming there are no practical issues with selecting either one (such as the amount of data available), which one would you choose for this task? Give **two** reasons why.

Transformers

Transformers don't have a recurrent computation, so the representations at each time step can directly attend to the representations at other time steps. As a result, it is more effective for modeling long-term dependencies because there is no vanishing gradients effect across time.

Because there is no recurrence, the representations at each time step can be computed in parallel

② **[1 pt]** Given that you have published 1000 columns at the Gazette, you have around 800 training examples that you can use for your system. Given the size of your dataset, do you think it would be helpful to pretrain your model on other text? Why or why not?

Yes, 800 columns would not be enough training data to learn a suitable generator.

③ **[1 pt]** Would you use a causal language modeling or masked language modeling training objective to train your model? Why?

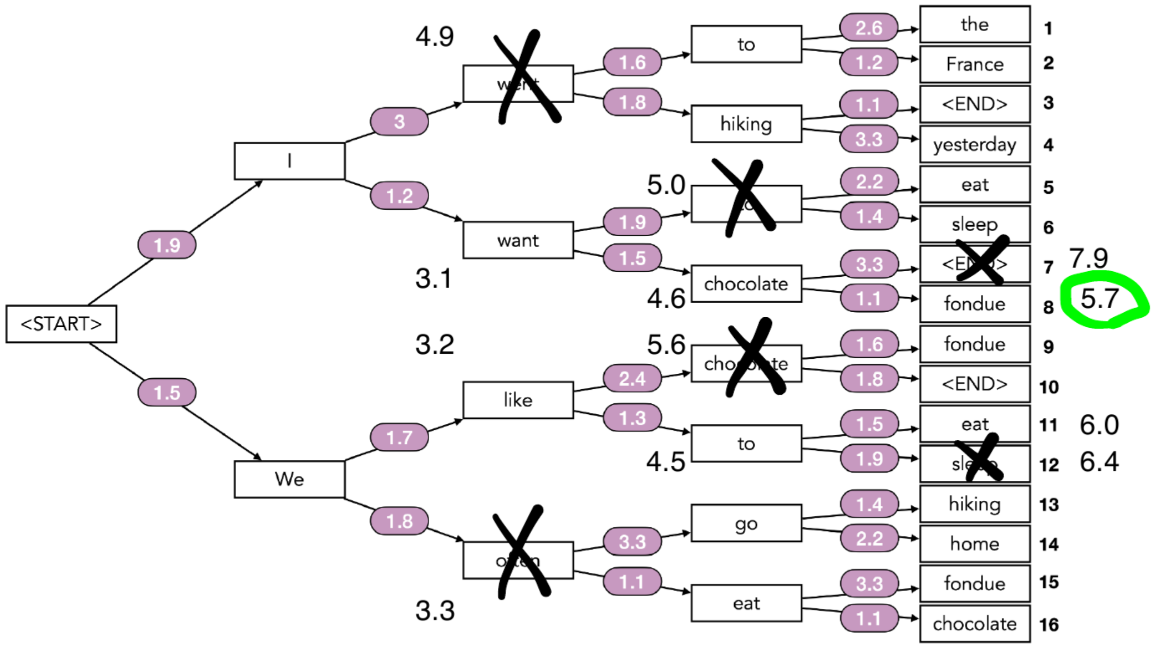
causal language modeling

learns to predict the next word, which you would need to generate a story.

④ **[1 pt]** You initialize your model with a vocabulary V with $|V|$ tokens. Given a vector of scores $S = [s_1, \ldots, s_i, \ldots, s_{|V|}]$ output by your model for each token in your vocabulary, write out the softmax function to convert score s_1 to a probability mass $P(s_1)$:

$$\frac{\exp(s_1)}{\sum_{i=1}^{|V|} \exp(s_i)}$$

⑤ **[7 pt]** Now that you've trained your model on your dataset, you can produce text from it. You pre-compute the step-by-step probability distributions over all tokens for four steps. Below, we show the top-2 highest probability tokens in these distributions at each step (along with their **negative log probability**):



For each of the following sub-questions a, b and c, provide your answer in a form similar to:
<START> I went hiking yesterday

a) **[1 pt]** What sequence would be produced using argmax decoding?

`<START> We like to eat`

$$1.5 + 1.7 + 1.3 + 1.5 = 6.0$$

b) **[4 pt]** What sequence would be produced using beam search with a beam size of 2? **Justify** your answer by *annotating* the above graph with choices and calculations.

`<START> I want chocolate fondue`

$$1.9 + 1.2 + 1.5 + 1.1 = 5.7$$

c) **[2 pt]** What is the optimal sequence?

`<START> We often eat chocolate`

$$1.5 + 1.8 + 1.1 + 1.1 = 5.5$$

⑥ **[4 pt]** You are given a probability distribution $P(y_t|y_0, \dots, y_{t-1})$ over 100 possible next tokens to generate by your model. The distribution has the following characteristics:

- 20% of the probability mass is on the most probable token;
- 10% of the probability mass is on each of the next 4 most probable tokens;
- 1% of the probability mass is on each of the next 20 most probable tokens;
- the remaining mass is uniformly distributed across the remaining 75 tokens.

a) **[2 pt]** In top-k sampling, if $k = 15$, how much probability mass will be included in the set of tokens you sample from?
Fully justify your answer.

first is 20%;

2 to 5 are 10% each thus 40% altogether;

6 to 15 are 1% each, thus 10% altogether;

the total is **70%**

b) **[2 pt]** In top-p sampling, if $p = 0.75$, how many tokens will be included in the set of tokens you sample from?
Fully justify your answer.

Following up on former question: we still lack 5%: 16 to 20 will provide it: **20 tokens** altogether.

⑦ **[4 pt]** The outputs of which decoding algorithm would be changed by increasing the temperature hyperparameter of the softmax calculation? Assume the seed of your random number generator remains the same. Select all that apply and **justify** your answer for each:

[no] Argmax decoding: the top token remains the same even though the temperature changes

[yes] Beam search decoding: distribution is changed by increase in temperature. changes in relative probabilities could change the optimal branches to keep in the search

[yes] Top-k sampling: distribution is changed by increase in temperature. While the top k tokens would be the same, their relative probability mass would be different and the random number generator would change the output

[yes] Top-p sampling: distribution is changed by increase in temperature, so the same random number generator would change the output

⑧ **[4 pt]** To evaluate your system, you decide to hold out some of the columns you have previously written and use them as an evaluation set. After generating new columns using the same titles as these held-out columns, you decide to evaluate their quality.

a) **[1 pt]** What would be an advantage of using a content overlap metric?

cheap to implement, fast to run

b) **[1 pt]** What would be a disadvantage of using a content overlap metric?

content overlap metrics do not measure more than the overlap of words in generated and reference sequences so words like synonyms would not be rewarded by similar phrasings of opposite semantic could be

c) **[1 pt]** What would be an advantage of using a model-based metric?

- representing words as embeddings allows finer-grained semantics beyond word overlap to be captured by the metric
- more correlated with human judgment

d) **[1 pt]** What would be a disadvantage of using a model-based metric?

- not interpretable
- requires training on a corpus of annotated scores

⑨ **[3 pt]** Your column generation system has become quite successful and you've managed to automate most of your job simply by typing your editor's title pitches into your model to produce your column every day. Two years later, during the COVID–25 pandemic, your editor proposes to use your system to generate an information sheet about the pandemic for anyone looking for information about symptoms, treatments, testing sites, medical professionals, etc. Given the similarity to a previous pandemic many years before, COVID–19, you train your model on all news articles published about COVID–19 between the years of 2019–2022. Then, you generate the information page from your trained model.

Give an example of a potential harm that your model could produce from the perspective of:

- leaking private information;
- disinformation;
- human interaction harms.

Leaking Private Information: Previous data could have mentioned names, addresses, titles, workplaces, of medical professionals during COVID-19. This information could be generated by the model if trained on this data

Disinformation: The model could generate fake content such as symptoms, potential treatments, addresses of known professional. Some of these may have been true for COVID-19, but now are false for COVID-25.

Human interaction harms: The model could generate text that suggests treatments to users. As the model is not a medical professional, these treatments could cause harm to the user if followed. The model could also give wrong addresses to testing sites, causing users to be harmed. Others are acceptable.

⑩ **[2 pt]** In your writings on the topic of medicine, you typically quoted two professionals, Dr. John Smith (around 60% of the time) and Dr. Virginia Jones (around 40% of the time). Would you expect the proportion of quotes by Dr. John Smith to be less than, greater than, or around 60% in the generated stories your model produces? Why?

More than 60%

Bias amplification

QUESTION II : Pulsed lasers [24 pt]

Consider the following sentence:

High-energy pulsed laser beams are used in soft-tissue surgery.

① **[1 pt]** Using a tokenizer that splits on whitespaces and punctuation (including hyphens (-)), what is the token sequence?

High, –, energy, pulsed, laser, beams, are, used, in, soft, –, tissue, surgery, .

(here keeping the punctuation signs, but they could also be removed).

② **[1 pt]** Using a 1-gram language model and the same tokenizer, what is the probability of the above sentence? Provide your answer as a formula, but clearly explaining each variable.

It's $\prod_{i=1}^{14} P(w_i)$ with the above fourteen tokens w_i (eleven, if punctuation has been removed).

③ **[2 pt]** Same question but using a 2-gram language model:

It's $P(\text{High}) \cdot \prod_{i=2}^{14} P(w_i|w_{i-1})$ with the same fourteen w_i as before.

④ **[4 pt]** Tokenization is now enhanced with Named Entity Recognition (NER) specialized on technical and medical terms.

a) **[1 pt]** How is your answer to question ② modified? **Fully justify** your answer.

b) **[3 pt]** What would be the advantage of doing so? What would be the major drawback? Justify your answers.

a) Assume that after NER, the tokenization is:
High-energy, pulsed, laser beams, are, used, in, soft–tissue, surgery, .
or even:
High-energy pulsed laser beams, are, used, in, soft–tissue surgery, .
then the formula will be the same but with these new, much less, tokens (nine in the first case and six in the second).

b) Doing so improves probability estimation and inference (provided that we have enough learning data), because these are not independent terms, thus the probabilities of the NERs are not (higher) the products of their terms (probabilistic independence).

Drawback: NER can be wrong at this stage. It's better not to take decision too early in the process: shipping all the alternatives to the next modules would be better.

Consider now the following toy learning corpus[1] of 59 tokens[2], out of a possible vocabulary of $N = 100$ different tokens:

Pulsed operation of lasers refers to any laser not classified as continuous wave, so that the optical power appears in pulses of some duration at some repetition rate. This

[1] [excerpt from https://en.wikipedia.org/wiki/Pulsed_laser]
[2] The same tokenizer was used, but *without* any NER.

encompasses a wide range of technologies addressing a number of different motivations. Some lasers are pulsed simply because they cannot be run in continuous wave mode.

⑤ **[3 pt]** Using a 2-gram language model, what are the values of the parameters corresponding to "*continuous wave*" and to "*pulsed laser*"...

a) [1 pt] ...using Maximum-Likelihood estimates?

b) [2 pt] ...using estimation smoothed by a Dirichlet prior with parameters all equal to 0.01?

Justify your answers.

a) $P(\text{continuous wave}) = \frac{2}{58}$ since "*continuous wave*" appears two times and that there are 58 bigrams in a 59 token corpus.

$P(\text{pulsed laser}) = 0$ since "*pulsed laser*" never occurs in the corpus.

b) $P(\text{continuous wave}) = \frac{2.01}{58 + 0.01 \times N^2} = \frac{2.01}{158}$ where N is the size of the (total possible) vocabulary.

$P(\text{pulsed laser}) = \frac{0.01}{58 + 0.01 \times N^2} = \frac{1}{15\,800}$

Consider now the following shorter phrase:

laser used for surgery process

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

- `N`: 0.1, `V`: 0.15, `Adj`: 0.2, `Prep`: 0.05, ...
-

laser:	`Adj`:	$4 \cdot 10^{-4}$,	`N`:	$5 \cdot 10^{-4}$
used:	`Adj`:	$8 \cdot 10^{-4}$,	`V`:	$6 \cdot 10^{-4}$
for:	`Prep`:	$9.5 \cdot 10^{-4}$		
surgery:	`N`:	$7.3 \cdot 10^{-4}$		
process:	`N`:	$7 \cdot 10^{-4}$,	`V`:	$5 \cdot 10^{-4}$

-

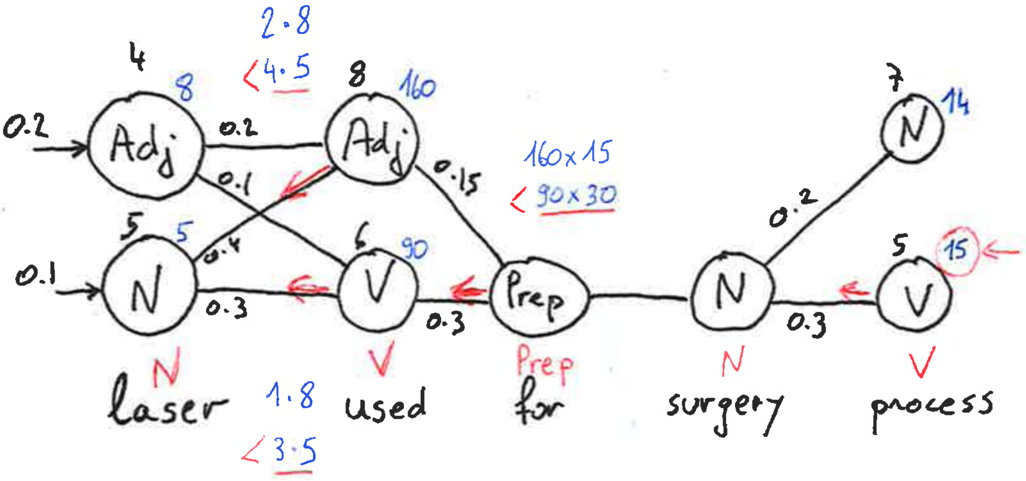
	`Adj`	`N`	`V`	`Prep`
`Adj`	0.2	0.5	0.1	0.15
`N`	0.4	0.2	0.3	0.1
`V`	0.1	0.4	0.15	0.3
`Prep`	0.02	0.45	0.51	0.01

⑥ **[7 pt]** What is the most probable sequence of tags for the above sentence?

Fully justify your answer.

First notice that the above table contains transition probabilities from row tag to column tag (look at the sum of the second (or third) column).

Use the Viterbi algorithm (rather than brute force):



The most probable sequence is then: `N V Prep N V`

Pay attention to use initial probabilites, as well as transition to `Prep`.

continues on back ☞

Finally, we'd like to do some sentence topic classification using a Naive-Bayes model.

Consider the following toy learning corpus, where each sentence has been assigned a topic, either "Medical" or "Computer":

- **Medical**: plastic surgery process initial consultation can be scheduled by sending an email to the administration.
- **Medical**: in the process, the laser beam comes into contact with soft tissues.
- **Medical**: laser eye surgery process reshapes parts of the cornea by removing tiny amount of tissues.
- **Computer**: the team behind the laser based quantum computer includes scientists from the US, Australia and Japan.
- **Computer**: the optical laser barcode scanner was plugged on the USB port.
- **Computer**: cdrom laser lens cleaning process starts with opening the tray.
- **Computer**: laser was a computer trademark

The parameters are learned using some appropriate additive smoothing with the same value for all parameters. In the above learning corpus, there are 42 token occurrences in "Medical" documents and 42 token occurrences in "Computer" documents (punctuation is ignored).

⑦ **[6 pt]** How would the following short sentence:

pulsed laser used for surgery process

be classified by this model?

Fully justify your answer and mathematically support your claim.

Answer: **Medical** (not hard to guess ;-))

Justification:

- Priors: **Medical**: $\frac{3}{7}$, **Computer**: $\frac{4}{7}$
- counts:

	laser	surgery	process
Medical:	2	2	3
Computer:	4	0	1

- inference (preprocessing removing "*used for*", or will have the same parameters anyway):

$$P(\text{input}, \textbf{Medical}) \propto \frac{3}{7} \times \frac{\alpha}{42 + N\alpha} \times \frac{2+\alpha}{42 + N\alpha} \times \frac{2+\alpha}{42 + N\alpha} \times \frac{3+\alpha}{42 + N\alpha}$$

$$P(\text{input}, \textbf{Computer}) \propto \frac{4}{7} \times \frac{\alpha}{42 + N\alpha} \times \frac{4+\alpha}{42 + N\alpha} \times \frac{\alpha}{42 + N\alpha} \times \frac{1+\alpha}{42 + N\alpha}$$

which, with decent α leads to approximation:

$$P(\text{input}, \textbf{Medical}) \propto\simeq 36$$

$$P(\text{input}, \mathbf{Computer}) \propto\simeq 16\alpha$$

The former being at least one order of magnitude bigger.

QUESTION III : Systems and queries [17 pt]

Consider the following evaluation of two IR systems made on a toy corpus of four queries, where ✔ denotes a retrieved relevant document and ✗ denotes a retrieved non-relevant document (the two systems may not retrieve the same document at each rank):

query q_1:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✔	✔	✗	✗	✗	✔	✔
system 2	✗	✔	✗	✔	✔	✔	✔	✔

query q_2:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✔	✔	✗	✗	✔	✔	✔
system 2	✗	✔	✔	✔	✔	✔	✔	✔

query q_3:

	rank							
	1	2	3	4	5	6	7	8
system 1	✔	✗	✔	✗	✔	✗	✔	✗
system 2	✔	✔	✗	✗	✔	✔	✗	✗

query q_4:

	rank							
	1	2	3	4	5	6	7	8
system 1	✗	✔	✔	✔	✗	✗	✔	✔
system 2	✔	✗	✔	✗	✔	✔	✗	✗

In the above results, we assume that, for each query, at least one of the two systems retrieved **all** the relevant documents; and that the missing relevant documents are never retrieved.

① **[1 pt]** For each of the four queries, what is the total number of relevant documents?

Number of relevant documents: q_1: 6, q_2: 7, q_3: 4, q_4: 5

② **[1 pt]** What is the *recall* of system 1 for query q_1?

Provide your answer in the form of a fraction and **justify** your answer.

$R = \frac{5}{6}$ as there are six relevant documents in the corpus for query q_1, among which system 1 only retrieves five.

③ **[2 pt]** What is P@6 for each of the two systems for query q_1?

Provide your answers in the form of fractions and **justify** your answers.

system 1: $P_6(q_1) = \frac{3}{6} = \frac{1}{2}$ since it retrieves 3 relevant document out of six

system 2: $P_6(q_1) = \frac{4}{6} = \frac{2}{3}$ similarly

④ **[4 pt]** What is the R-Precision for each of the **two** systems?

Provide your answers as simple arithmetic expressions involving only a few fractions and **justify** your answers.

system 1: $P_6(q_1) = \frac{1}{2}$, $P_7(q_2) = \frac{5}{7}$, $P_4(q_3) = \frac{1}{2}$, $P_5(q_4) = \frac{3}{5}$

$$\text{R-Prec} = \frac{1}{4}\left(\frac{1}{2} + \frac{5}{7} + \frac{1}{2} + \frac{3}{5}\right) = \frac{1}{4} \cdot \frac{81}{35} = \frac{81}{140}$$

system 2: $P_6(q_1) = \frac{2}{3}$, $P_7(q_2) = \frac{6}{7}$, $P_4(q_3) = \frac{1}{2}$, $P_5(q_4) = \frac{3}{5}$

$$\text{R-Prec} = \frac{1}{4}\left(\frac{2}{3} + \frac{6}{7} + \frac{1}{2} + \frac{3}{5}\right) = \frac{1}{4} \cdot \frac{551}{210} = \frac{551}{840}$$

⑤ **[5 pt]** What is the MAP for system 1?

Provide your answers in the form: $\frac{1}{n}(A_1 + A_2 + ...)$ by providing the value of n and expressing each of the A_i as a simple arithmetic expression involving only a few fractions.

Then **justify** your answers.

$$A_1 = \text{AvgP}(q_1) = \frac{1}{6}\left(1+1+1+\frac{4}{7}+\frac{5}{8}\right)$$

$$A_2 = \text{AvgP}(q_2) = \frac{1}{7}\left(1+1+1+\frac{4}{6}+\frac{5}{7}+\frac{6}{8}\right)$$

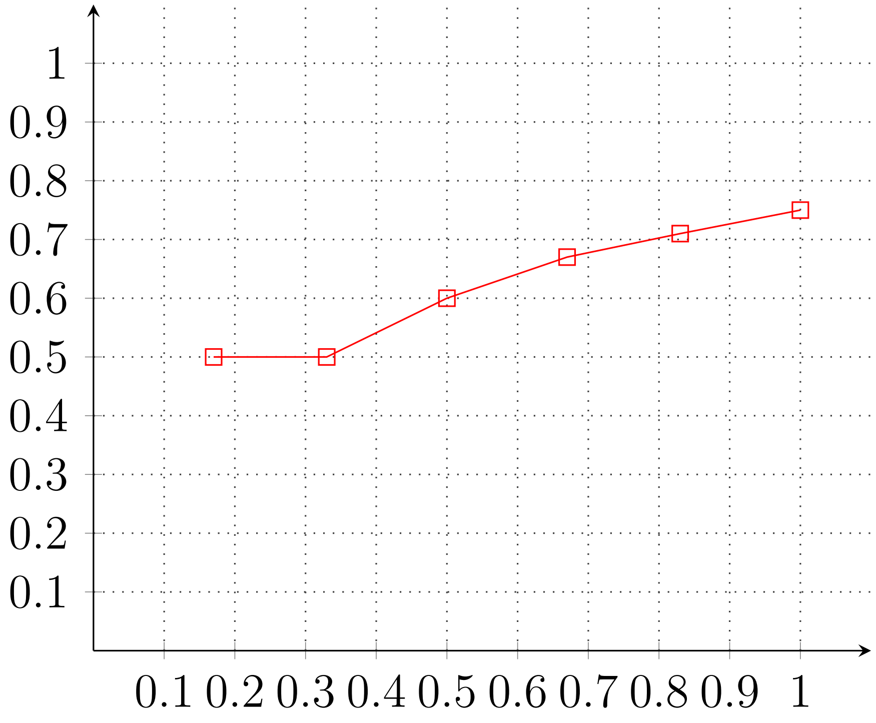
$$A_3 = \text{AvgP}(q_3) = \frac{1}{4}\left(1+\frac{2}{3}+\frac{3}{5}+\frac{4}{7}\right)$$

$$A_4 = \text{AvgP}(q_4) = \frac{1}{5}\left(\frac{1}{2}+\frac{2}{3}+\frac{3}{4}+\frac{4}{7}+\frac{5}{8}\right)$$

and $n = 4$.

⑥ **[4 pt]** Draw the P-R curve for system **2** and query q_1 only (using P@k).

Justify your answer and **explain** what each of the two axis represents.



A few numerical approximations:

$\frac{1}{6} \simeq 0.15$	$\frac{5}{6} \simeq 0.83$
$\frac{1}{7} \simeq 0.14$	$\frac{3}{7} \simeq 0.42$
$\frac{5}{7} \simeq 0.71$	$\frac{6}{7} \simeq 0.85$
$\frac{1}{8} \simeq 0.12$	$\frac{3}{8} \simeq 0.37$
$\frac{5}{8} \simeq 0.62$	$\frac{7}{8} \simeq 0.87$

x axis is recall, y axis is precision.

Points are:

$$(\frac{1}{6},\frac{1}{2}) \quad (\frac{2}{6},\frac{1}{2}) \quad (\frac{3}{6},\frac{3}{5}) \quad (\frac{4}{6},\frac{4}{6}) \quad (\frac{5}{6},\frac{5}{7}) \quad (1,\frac{6}{8})$$

one point for each new relevant document retrieved (following system 2 ranks); the recall is then $i/6$ for i from 1 to 6.

QUESTION IV : About planes and balloons [9 pt]

You have been provided with the following definitions for the possible meanings of the words "balloon" and "plane":

```
balloon:
- meaning 1:
  balloon --(hyponym)--> inflatable

- meaning 2:
  balloon --(hyponym)--> transport
```

```
plane:
- meaning 1:
  plane --(hyponym)--> transport
  plane --(holonym)--> wing

- meaning 2:
  plane --(hyponym)--> surface
```

① **[2 pt]** How would you express the provided four meanings in plain English?

```
balloon
1: a kind of inflatable;
2: a kind of transport.

plane
1: a kind of transport with wings;
2: a kind of surface.
```

Definition containing any additional information are not considered as acceptable, as the additional information is not mentioned in the definition.

② **[1 pt]** What type of approach has been used to produce this type of semantic representations? What principle does it rely on?

The approach used is based on (a limited set of) semantic relations (hyponymy, holonymy) and relies on the Aristotelian "Genus-Differentia" principle.

③ **[2 pt]** Are the provided meaning representations formally correct? **Justify** your answer:

No. The provided representations are not formally correct, because semantic relations connect word meanings (usually represented as word_1, word_2, ...) and not words.

continues on back ☞

Assume that you are provided with a reference corpus consisting of a large amount of word definitions similar to the ones given above and that you are asked to design an evaluation metric aiming at assessing the quality of comparable definitions produced (for the same words) by various automated systems.

④ **[4 pt]** Do you think that an approach based on some kind of n-ary classification would be suitable? **Fully justify** you answer.

An approach based on some kind of n-ary classification would not be suitable for the following reasons:

1. 1. Whatever the metric that may be considered, its exploitability is likely to be very limited due to the inherent low quality of the reference resulting from the low inter-annotator agreement that is usually observed for many words when lexicographers are requested to produce definitions for them; in addition, there will be no guarantee that different lexicographers produce the same number of meanings for a given word, which would raise the additional issue of deciding how to align the meanings produced by an automated system with the ones present in the reference;
2. 2. If a binary classification should be considered, i.e. a definition would be tagged as either "correct" or "incorrect" (wrt. the provided reference definition), deciding about "correctness" would probably be difficult to automate, as it would, either be too strict (if an exact match would be requested), or hard to define properly (if some kind of partial match would be considered);
3. If a n-ary ($n > 2$) classification should be considered, the very high number of combination resulting from the association of several (`meaning_1`, `semantic_relation`, `meaning_2`) triples present in any definition would make the number n of different tags to consider too large for being exploitable in practice.

In short, using classes is not very well suited to capture the graph structure of the definitions.

QUESTION V : Conjugation [22 pt]

The goal of this question is to illustrate how to use transducers to implement a simplified version of the conjugation of English verbs. We will restrict to the conjugated forms corresponding to the indicative mode and the present tense. Formally, this can be modeled as defining a transducer able to recognize associations such as:

`make+V+IndPres+1s`	make
`make+V+IndPres+2s`	make
`make+V+IndPres+3s`	makes
`make+V+IndPres+1p`	make
`make+V+IndPres+2p`	make
`make+V+IndPres+3p`	make

where “`V`” identifies the grammatical category “Verb”, “`IndPres`” indicates that we are dealing with the Indicative mode and the Present tense, and “`1s`”, “`2s`”, “`3s`” (resp. “`1p`”, “`2p`”, “`3p`”) refer to the first, second, and third person singular (resp. plural).

① **[1 pt]** In the above table, what do the strings in the first and the second column correspond to?

he strings in the first column are canonical representations, while the strings in the second column are surface forms.

② **[1 pt]** Provide a detailed explanation of the interpretation of the strings present in the first column.

The typical format of a canonical representations is:

`Lemma+GrammaticalCategory+MorphoSyntacticFeature1+MorphoSyntacticFeature2+...`

Thus a canonical representation such as “`make+V+IndPres+1s`” should be interpreted as identifying the conjugated form (i.e. a verbal form, thus the grammatical category “`V`”) of the verb “to make” (thus the lemma “`make`”) at the 1st person singular (thus the morphosyntactic feature “`1s`”) of the present tense of the indicative mode (thus the morphosyntactic feature “`IndPres`”).

The idea is to build a transducer corresponding to the composition of three transducers:

- a transducer T_1 that defines the morphological paradigm, i.e. identifies the various cases to consider for conjugating a regular verb;
- a transducer T_2 that implements the identified cases in the form of transformation rules to be applied for the considered morphological paradigm;
- a transducer T_3 that handles all the exceptions to be implemented.

③ **[1 pt]** What is the number N of distinct cases to consider to conjugate *most* English verbs in present indicative? **Justify** your answer.

$N = 2$, because for most of the English verbs, there are only 2 distinct conjugated forms for the present indicative:

- the form corresponding to the 3rd person singular;
- the form corresponding to all the other [persons, number] combinations.

④ **[2 pt]** Describe in plain English the transformation rule associated with each of the N cases in the "present indicative" paradigm.

Rule 1: for the 3rd person singular, add a final "s" to the root;

Rule 2: for all other cases, leave the root unchanged.

⑤ **[2 pt]** Is the provided number N valid for all English verbs (indicative mode, present tense)? If yes, explain why; if not, provide a simple counter-example.

No. The $N = 2$ number is valid for most of the English verbs, but there are exceptions, such as the verb "to be", which corresponds to 3 distinct forms ("am", "are", and "is"), or modals, such as "can", which correspond to only 1 distinct form ("can").

⑥ **[2 pt]** Indicate in the table below the associations the transducer T_1 should recognize for the present indicative paradigm of the verb "to make" (each of the N cases mentioned in question ④ should be identified by a number between 1 and N; you can leave empty rows at the end of the table if necessary).

make+V+IndPres+1s	make+2
make+V+IndPres+2s	make+2
make+V+IndPres+3s	make+1
make+V+IndPres+1p	make+2
make+V+IndPres+2p	make+2
make+V+IndPres+3p	make+2

⑦ **[2 pt]** Provide a formal definition for transducer T_1:

The transducer T1 is built by using the standard operators (concatenation, disjunction and cross-product) and regular expressions available for the transducers.

For instance:

```
T1 = ([a-z]+)
     ((\+V\+IndPres\+) x (\+))
     (((([12]s)
        | ([123]p)) x (2))
      | ((3s) x (1))
     )
```

⑧ **[3 pt]** Provide a formal definition of the transducer T_2 that implements the rule(s) identified in former question ④. **Fully justify** your answer.

For instance:

```
T2 = ([a-z]+)
     (((\+1) x (Xs))
      | ((\+2) x (EPSILON)))
```

The rule for case 1 is implemented by replacing “+1” by “Xs”, i.e. adding to a final “s” to the root preceded by a “trace” X;

The rule for case 2 is implemented by removing “+2” (replace with the empty character), i.e. leaving the root unchanged;

⑨ **[1 pt]** What is the number M of (distinct) associations to be recognized by the transducer T_2 for any verb with a “present indicative” paradigm corresponding to N cases?

$M = N = 2$

⑩ **[2 pt]** Indicate in the table below the associations the transducer T_2 should recognize to process the following three verbs: “to do”, “to make” and “to try” (you can leave empty rows at the end of the table if necessary, but put the associations in *alphabetic order*).

do+1	doXs
do+2	do
make+1	makeXs
make+2	make
try+1	tryXs
try+2	try

⑪ **[1 pt]** Provide a formal definition for the transducer T_3 that should be used *if there would be no exceptions*:

For instance:

```
T3 = ([a-z]+) ((Xs) x (s))
```

⑫ **[1 pt]** Indicate in the table below the strings that should be associated by the transducer T_3 defined in previous question ⑪ to the outputs of T_2 indicated in the table of question ⑩, and, for each of them, indicate whether they are correct or not by writing “OK” or “notOK” in the second column (you can leave empty rows at the end of the table if necessary).

dos	notOK
do	OK
makes	OK
make	OK
trys	notOK
try	OK

⑬ **[3 pt]** How should T_3 be modified to process correctly the cases identified as “notOK” in the table in question ⑫? Provide an updated formal definition for T_3.

For instance:

```
T3 = ([a-z]+)
     (   ((oXs) x (oes))
      |  ((yXs) x (ies))
      |  ([^oy]((Xs) x (s)))
      |  (EPSILON)
     )
```

QUESTION VI : Ambiguities [14 pt]

Consider the (toy) grammar G consisting of the following rules:

```
R1: S --> NP VP
R2: NP --> NN
R3: NP --> Det NN
R4: NN --> N
R5: NN --> NN NN
R6: NN --> NN PNP
R7: PNP --> Prep NP
R8: VP --> V
R9: VP --> Adv V
```

① **[2 pt]** Precisely define the type of grammar G is corresponding to (for that, consider at least the following aspects: dependency-based vs. constituency-based, position in the Chomsky hierarchy, and CNF); **justify** your answer for each of the aspects you will be mentioning.

The grammar G is:

- a constituency-based (because it consists of rewriting rules);
- context-free grammar (because of the format of the rules: only one an exactly one terminal on the left-hand side);
- in extended Chomsky Normal Form (because of the format of the rules: no more than two terms on the right-hand side).

② **[1 pt]** What type of rules does the provided grammar G consist of?
What type of rules should G be complemented with to be exploitable in practice?
What is the format of these missing rules?

G consists of syntactic rules.

G should be complemented with lexical rules with the following format: `T -> w`, where `T` is a pre-terminal (i.e. a Part-of-Speech tag) and `w` is a terminal (i.e. a word).

③ **[1 pt]** What is the number N of additional rules that should be added to G to make it applicable to any sequence of words from a set of $10\,000$ distinct words with an average syntactic ambiguity of 1.5? Justify your answer.

$N = 15\,000$

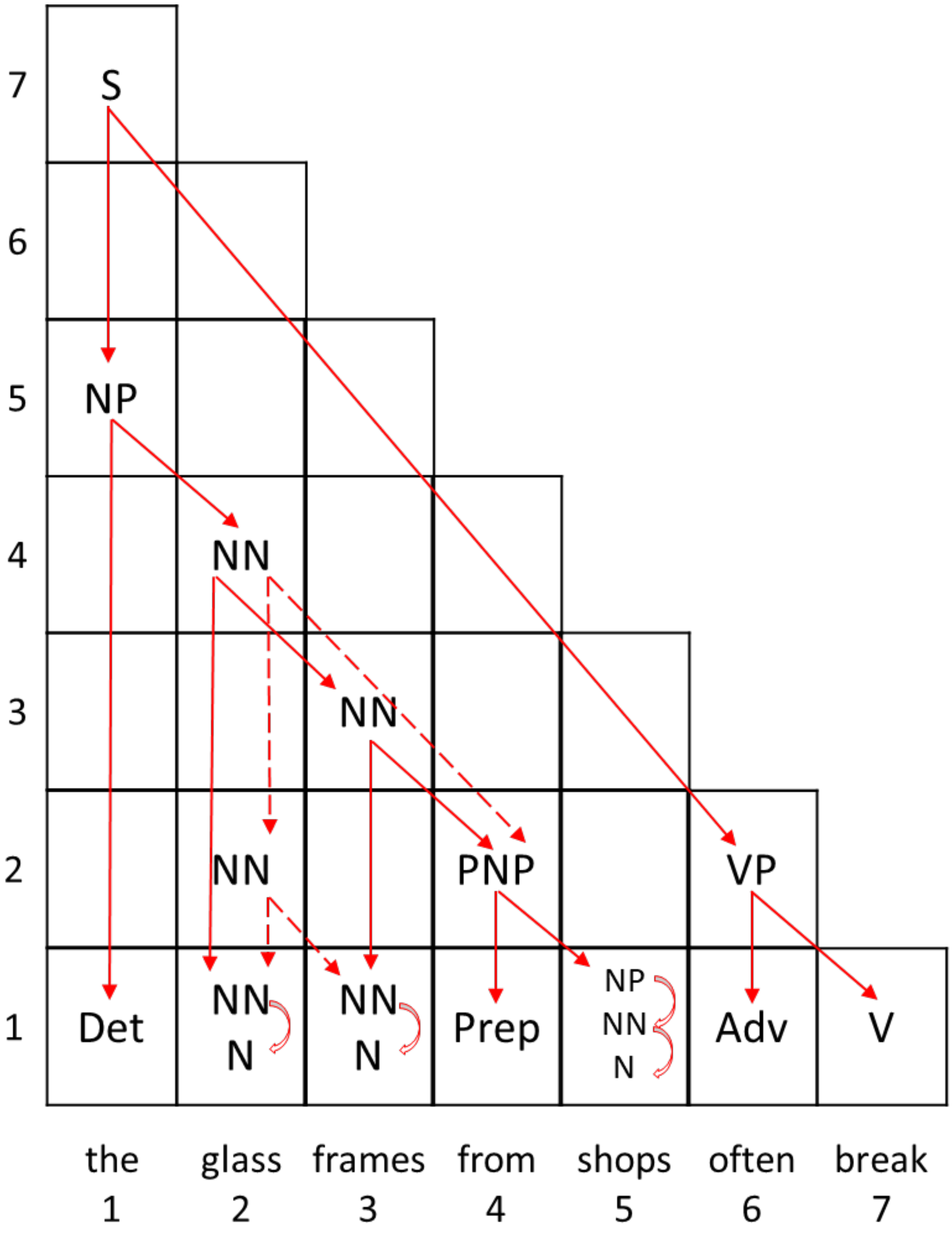
We need a lexical rule for each of the possible (PoS-tag, surface form) associations; there are 10’000 distinct words with an average syntactic ambiguity of 1.5, thus a total of 10’000*1.5 = 15’000 possible associations and therefore 15’000 lexical rules to be added.

④ **[6 pt]** Indicate the number K of parse trees produced by G for the sequence:

the glass frames from shops often break

and describe the identified parse tree(s) in the form of a CYK chart where *only* the non-terminals and the links *required* for retrieving *full* parse tree(s) are represented:

$K = 2$ possible parse trees described in the CYK table below:



⑤ **[2 pt]** Indicate what type of constraints are (resp. are not) taken into account by the grammar G, and, for each constraint type mentioned, provide illustrative examples.

The positional constraints (word order) are taken into account by G; for example, “the frames break” is accepted by G, as it should, and “frames the break” is not, as it should;

The selectional constraints (agreements) are not taken into account by G; for example, “the frames break” is accepted by G, as it should, but “the frame break” is also accepted by G, while it should not.

⑥ **[2 pt]** In how many rules should the 9 rules provided for G be expanded into to cope with simple number agreements? **Justify** your answer.

The provided 9 rules could be expanded as follows to take simple number agreement into account:

```
R1.1: S --> NPs VPs
R1.2: S --> NPp VPp
R2.1: NPs --> NNs
R2.2: NPp --> NNp
R3.1: NPs --> Dets NNs
R3.2: NPp --> Detp NNp
R4.1: NNs --> Ns
R4.2: NNp --> Np
R5.1: NNs --> NNs NNs
R5.2: NNp --> NNs NNp
R5.3: NNs --> NNp NNs
R5.4: NNp --> NNp NNp
R6.1: NNs --> NNs PNP
R6.2: NNp --> NNp PNP
R7.1: PNP --> Prep NPs
R7.2: PNP --> Prep NPp
R8.1: VPs --> Vs
R8.2: VPp --> Vp
R9: VPs --> Adv Vs
R9: VPs --> Adv Vs
```

thus resulting in a set of 20 syntactic rules.

Note that rule R5 may be expanded in only 2 rules instead of 4 if the assumption that in nominal compounds corresponding to a sequence of several (e.g. “satellite antenna frames”), all the nouns but the last one must be singular:

```
R5.1: NNs --> NNs NNs
R5.2: NNp --> NNs NNp
```