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 (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

# INTRODUCTION TO NATURAL LANGUAGE PROCESSING

## Solution and scoring scale to Fall 2020, Graded Quiz #4

**NAME:** **SCIPER:**

## Instructions:

You have 45 minutes (8:15–9:00) for this quiz, which consists of several questions of different weights. For each question the corresponding points are indicated and the total number of points is 30.
[...]

The solution comes in red. And some comments in blue. And grading scale in green.

## QUESTION I **[5 pt]**

Consider the following CF grammar G1

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

(where `Det`, `N`, `Prep` and `V` are the only pre-terminals),
complemented by an adequate lexicon L1.

① **[1 pt]** If the sequence $(p_1, p_2, ..., p_8)$ represents a set of probabilistic coefficients for the syntactic rules in G1 ($p_i$ being associated to $R_i$), indicate which of the following choices correspond to a valid probabilistic extension for the grammar G1:

(Select one or several answers. Penalty for wrong ticks.)

[50%] $(1.00, 0.00, 1.00, 0.00, 1.00, 0.00, 1.00, 0.00)$

[-100%] $(0.55, 0.45, 0.60, 0.10, 0.15, 0.75, 0.50, 0.50)$

[50%] $(0.35, 0.65, 1.00, 0.30, 0.25, 0.45, 0.25, 0.75)$

[-100%] I cannot answer because it also depends on the probabilistic coefficients associated to the lexical rules.

[-100%] None of the other proposed answers.

② **[1 pts]** Assume that the grammar G1 has been associated with a valid choice of probabilistic coefficients, but then needs to be converted into an *equivalent* SCFG in extended Chomsky Normal form.

Is it possible to derived the stochastic coefficients of the grammar resulting from the conversion from the ones of G1?

(Select only one answer.)

[✔] Yes.

[ ] No.

[ ] It depends on how the conversion has been done.

③ **[3 pts]** Consider a SCFG G2 consisting of the same syntactic rules as the ones provided for grammar G1, but, this time, complemented by the following lexicon L3:

```
ate: V
cat: N
cheese: N
mouse: N
the: Det
with: Prep
```

Indicate which of the following constraints should be enforced for the stochastic grammar G2 so that the most-probable parse is making use of the rule

```
NP -> Det N PNP
```

when parsing the sequence

*the cat ate the mouse with cheese*

(Select one or several answers. Penalty for wrong ticks.)

[-100%] $p_1 \cdot p_6 < p_2 \cdot p_5$

[50%] $p_2 < p_1 \cdot p_6$

The actual condition is $p_2 \cdot p_5 < p_1 \cdot p_6$, which is true if $p_2 < p_1 \cdot p_6$ since $p_5 \leq 1$.

[50%] $p_1 \cdot (1 - p_4) > p_5$

The actual condition is $p_2 \cdot p_5 < p_1 \cdot p_6$, which is equivalent to $(1 - p_1) \cdot p_5 < p_1 \cdot (1 - p_4 - p_5)$.

[-100%] I cannot answer because it depends on the stochastic coefficients associated to the lexical rules.

## QUESTION II [2 pt]

Consider the following CYK chart and SCFG excerpt:

| | | |
|---|---|---|
| A: $x$ | | |
| B: 0.1 | C: $y$ | |
| D: 0.5 | E: 0.8<br>F: 0.9 | G: 0.2<br>H: 0.3 |

```
A  -->  B  H  (0.4)
A  -->  D  C  (0.6)

C  -->  E  G  (0.8)
C  -->  F  H  (0.2)
```

The notation "X: $p$" in a cell represents the information $p$ associated to non-terminal X so as to compute the most-probable parse.

What is the *numerical* value of $x$?

$x = 0.6 \times 0.5 \times y = 0.6 \times 0.5 \times 0.128 = 6 \times 64 \times 10^{-4} = 0.0384$, since:

$$0.8 \times 0.8 \times 0.2 > 0.2 \times 0.9 \times 0.3$$

thus $y = 0.8 \times 0.8 \times 0.2 = 0.128$

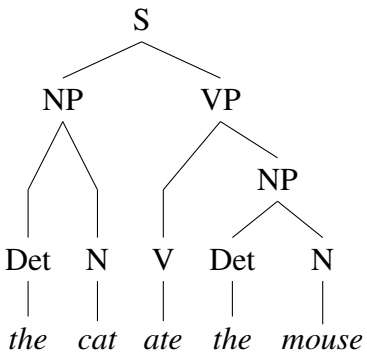
and

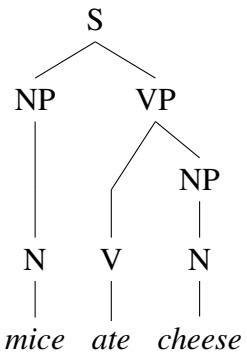
$$0.6 \times 0.5 \times y > 0.4 \times 0.1 \times 0.3$$

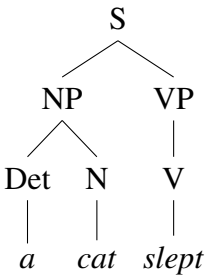
**continues on back ☞**

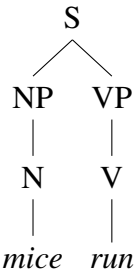
## QUESTION III [4 pt]

Assume that the following tree bank is available:




and let us call G3 the CFG grammar that can be extracted from this tree bank.

① **[2.5 pt]** Indicate both how many rules in total and how many *syntactic* rules are contained in G3. Format your answer as "$n_1$, $n_2$", where $n_1$ is the total number of rules and $n_2$ is the number of *syntactic* rules.

14, 5

② **[1.5 pt]** If the available tree bank is used to derive MLE estimates for the stochastic coefficients to be associated with the rules in G2, what is the estimate produced for the rule "`NP -> Det N`"?

3/6: 3 trees out of all the 6 `NP`-rooted trees.

## QUESTION IV [4 pt]

A query q has been submitted to two distinct Information Retrieval engines operating on the same document collection containing 1'000 documents, with 50 documents being truly relevant for q.

The following result lists have been produced by the two IR engines, S1 and S2 respectively:

| S1: | S2: |
|---|---|
| d1 | d'1 (⋆) |
| d2 (⋆) | d'2 (⋆) |
| d3 (⋆) | d'3 |
| d4 | d'4 |
| d5 (⋆) | d'5 |

In these result lists, the stars (⋆) identify the truly relevant documents.
By convention, we consider that any non retrieved document has been retrieved at rank 6.

① **[3 pt]** If Average Precision is used as evaluation metric, which of the two IR engines is performing better for the query q?

(Select only one answer.)

[ ] S1
[✔] S2
[ ] Both engines perform equally.
[ ] This evaluation metric cannot be computed.

$$\begin{aligned}
\text{AvgP}(S_1, q) &= \frac{1}{50}\ (\text{P@2}(S_1, q) + \text{P@3}(S_1, q) + \text{P@5}(S_1, q) + 47 \times \text{P@6}(S_1, q)) \\
&= \frac{1}{50}\ \left(\frac{1}{2} + \frac{2}{3} + \frac{3}{5} + 47 \times \frac{50}{1000}\right) \\
&\simeq \frac{1}{50}\ (0.50 + 0.67 + 0.60) + \frac{47}{1000} = \frac{1.77}{50} + \frac{47}{1000} \\
\text{AvgP}(S_2, q) &= \frac{1}{50}\ (\text{P@1}(S_2, q) + \text{P@2}(S_2, q) + 48 \times \text{P@6}(S_2, q)) \\
&= \frac{1}{50}\ \left(1 + 1 + 48 \times \frac{50}{1000}\right) \\
&\simeq \frac{1}{50}\ \left(1 + 1 + \frac{50}{1000}\right) + \frac{47}{1000} = \frac{2.05}{50} + \frac{47}{1000}
\end{aligned}$$

② **[1 pt]** Same question as in Q6, but with R-Precision used as evaluation metric for the single query q.

[ ] S1
[ ] S2
[ ] Both engines perform equally.
[✔] This evaluation metric cannot be computed.

Here R-Precision is equal to P@50(q), which cannot be computed because we can't identify the 45 first documents retrieved after rank 5.

Mathematically speaking, we have:

$$\frac{3}{50} \leq \text{R-Prec}(S_1, q) \leq \frac{48}{50}$$

and

$$\frac{2}{50} \leq \text{R-Prec}(S_2, q) \leq \frac{47}{50}$$

From which we cannot conclude.

## QUESTION V [3 pt]

Consider an IR engine, which uses an indexing mechanism implementing the following 3 consecutive filters:

1. a morpho-syntactic filter that restrict indexing term candidates to only nouns, and reduces them to their root form;
2. a frequencial filter parameterized with $f_{\text{min}} = 0.06$ (resp. $f_{\text{max}} = 0.20$) as lower (resp. upper) cut-off value, expressed as *relative* frequencies;
3. a stop word filter using the following stop list: *a*, *in*, *mouse*, *the* .

and the following document d:

*Cats are the worst enemies of rodents. After all, a cat is a cat: as soon as it can, it rushes into the bushes with only one target in mind: mice, mice and mice!*
*Naturally, the cats of houses are less frightening, as for them croquette loaded dressers have replaced prey hiding bushes. Cat's life in the house is easy!...*

What is the multi-set resulting from the indexing of document d by the above described IR engine?

Format your answer as an *alphabetically ordered* list of the form: "lemma1(tf1), lemma2(tf2), ...", where tfi is the term frequency of indexing term i.

For instance: dog(2), frog(3), zebra(1)

**Answer:** bush(2), house(2)

After step 1:

*cat enemy rodent cat cat bush target mind mouse mouse mouse house cat croquette dresser prey bush house cat life*

sort+frequencies (total = 20):

bush(2), cat(5), croquette(1), dresser(1), enemy(1), house(2), life(1), mind(1), mouse(3), prey(1), rodent(1), target(1)

frequency filtering ($f_{\text{min}} = 0.06 \times 20 = 1.2$, $f_{\text{max}} = 0.20 \times 20 = 4$:

bush(2), house(2), mouse(3)

stop-list filtering:

bush(2), house(2)

## QUESTION VI [4 pt]

Consider an IR system using a Vector Space model with Okapi BM25 as the weighting scheme (with $k = 1.5$ and $b = 0.75$) and operating on a document collection that contains:

- a document $d_1$, and
- and a document $d_3$ corresponding to the concatenation of 3 copies of $d_1$.

Indicate which of the following statements are true, where $\langle d \rangle$ stands for the vector representing document $d$:

(Select one or several answers. Penalty for wrong ticks.)

[- 33%] The cosine similarity between $\langle d_3 \rangle$ and $\langle d_1 \rangle$ is equal to 1.

[50%] Each component of $\langle d_3 \rangle$ is strictly larger than the corresponding one in $\langle d_1 \rangle$.

[-50%] Each component of $\langle d_3 \rangle$ is strictly smaller than the corresponding one in $\langle d_1 \rangle$.

[50%] Indexing terms with small term frequency are favored in $\langle d_3 \rangle$ (w.r.t. $\langle d_1 \rangle$).

[-50%] Indexing terms with large term frequency are favored in $\langle d_3 \rangle$ (w.r.t. $\langle d_1 \rangle$).

$$w(t,d) = \frac{k+1}{\text{tf}(t,d) + \frac{k}{\text{avdl}}|d| + k\,(1-b)}\ \text{tf}(t,d)\ \text{idf}(t)$$

The only terms that vary between $\langle d_1 \rangle$ and $\langle d_3 \rangle$ are:

$$\text{tf}(t,d_3) = 3\ \text{tf}(t,d_1)$$
$$|d_3| = 3\ |d_1|$$

Thus

$$\frac{w(t,d_3)}{w(t,d_1)} = \frac{\text{tf}(t,d_1) + A + B}{\text{tf}(t,d_1) + A + \frac{1}{3}B} > 1$$

with $A = \frac{k}{\text{avdl}}$ and $B = k\,(1-b)$

This ratio is moreover decreasing with $\text{tf}(t,d_1)$ (it's derivative is $-\frac{2B}{3\,(\cdots)^2}$), thus indexing terms with small term frequency are favored in $\langle d_3 \rangle$.

Notice that then $\langle d_3 \rangle$ cannot be colinear to $\langle d_1 \rangle$ in general (some component are more favoured than others), thus the cosine is not 1 in general.

## QUESTION VII [4 pt]

Consider the last 3 steps of a dendrogram clustering using complete linkage, where clusters A, B, C and D have the following distances:

| | B | C | D |
|---|---|---|---|
| A | 0.1 | 0.3 | 0.6 |
| B | | 0.4 | 0.2 |
| C | | | 0.5 |

① **[2 pt]** What are the last two (top levels) clusters? Provide your answer with only one comma (,) separating the two clusters. For instance if the last two clusters are A and B on one side and C and D on the other, answer: AB, CD; if the last two clusters are A on one side and B, C and D on the other, answer: A, BCD.

**Answer** ABC, D.

We indeed first regroup A and B since they are the closest.

We thus have to compute $d(AB, C)$ and $d(AB, D)$, which, with complete linkage are: $d(AB, C) = 0.4$ and $d(AB, D) = 0.6$.

Thus the next two groups which are joined are $AB$ and $C$ (smallest distance, $0.4$).

And $d(ABC, D) = 0.6$ (complete linkage).

② **[2 pt]** What is the distance between these last two clusters?

$0.6$

## QUESTION VIII [4 pt]

Consider a Naive Bayes classifier for ads with 3 classes: vehicles, real estate and jobs.

① **[1 pt]** What parameter(s) would such a classifier use to classify an ad (of at least two words) containing the word "*salary*"?

(Select one or several answers. Penalty for wrong ticks.)

[100%] $P(\text{"}salary\text{"}|\text{vehicles})$

[-100%] $P(\text{job}|\text{"}salary\text{"})$

[-100%] $P(\text{real estate}, \text{"}salary\text{"})$

[-100%] None of the other proposed answers.

② **[2 pt]** Knowing that we have 25% of vehicles ads and 40% of real estate ads, how would the following (fake) add be classified

*hello world*

if the parameters (as questioned in former question) are (here expressed up to some unique same multiplicative constant):

| | vehicles | real estate | jobs |
|---|---|---|---|
| hello | 6 | 2 | 4 |
| world | 5 | 8 | 7 |

**Answer:** jobs

vehicles: $\propto 0.25 \times 6 \times 5 = 7.5$

real estate: $\propto 0.40 \times 2 \times 8 = 6.4$

jobs: $\propto 0.35 \times 4 \times 7 = 9.8$