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

School of Computer and Communication Sciences
CS-431: Introduction to Natural Language Processing
Chappelier, J.-C. & Rajman, M.

CS–431: Introduction to Natural Language Processing

Exercises

(version 202501–3)

Contents

1 NLP levels

Exercise I.1

A company active in automatic recognition of hand-written documents needs to improve the quality of their recognizer. This recognizer produces sets of sequences of correct English words, but some of the produced sequences do not make any sense. For instance the processing of a given hand-written input can produce a set of transcriptions like: "*A was salmon outer the does*", "*It was a afternoon nice sunny*", and "*I Thomas at mice not the spoon*".

What is wrong with such sentences? NLP techniques of what level might allow the system to select the correct one(s)? What would be the required resources?

2 Evaluation

Exercise II.1

① Give some arguments justifying why evaluation is especially important for NLP. In particular, explain the role of evaluation when a corpus-based approach is used.

② Many general evaluation metrics can be considered for various NLP tasks. The simplest one is accuracy.

Give several examples of NLP tasks for which accuracy can be used as an evaluation metric. Justify why.

In general, what property(ies) must an NLP task satisfy in order to be evaluable through accuracy?

③ Consider a Part-of-Speech tagger[1] producing the following output:

```
The/Determiner program/Noun can/Noun deal/Noun with/Preposition three/Number
types/Verb of/Preposition inputs/Noun ./Punctuation
```

(using your own knowledge of general English,) Compute the accuracy of the tagger.

What do you think of the performance of this system with respect to the State of the Art? Is this conclusion reliable?

④ What is the formal relation between accuracy and the error rate? In which case would you recommend to use the one or the other?

⑤ Consider the following "breaking news scanning system":
A company receives a continuous stream of information messages (newswires); each time a new message arrives, its average textual similarity score with respect to the stored collection of previously received messages is computed. If this average similarity is below a given threshold, the message is considered "breaking news" and is automatically distributed to the company personnel.

The company has carried out an evaluation of the system in place, which produced the following average figures:

- one message out of 1000 is considered to be "breaking news" by the system;
- 30% of the claimed "breaking news" messages are evaluated as not new by human judges;
- the system is missing one truly "breaking news" message every 1000 messages processed.

Use the provided figures to compute the accuracy of the system.

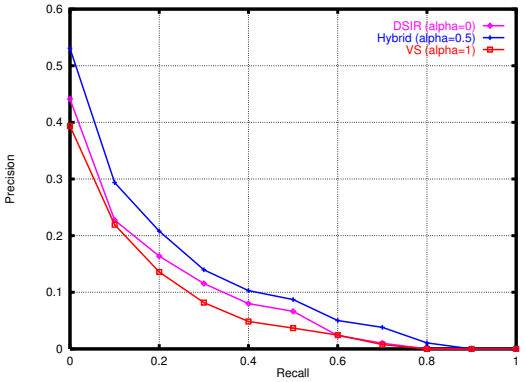
Is accuracy a good metric in this case? Justify your answer, and, possibly, propose some alternative performance score(s) and compute the corresponding value(s).

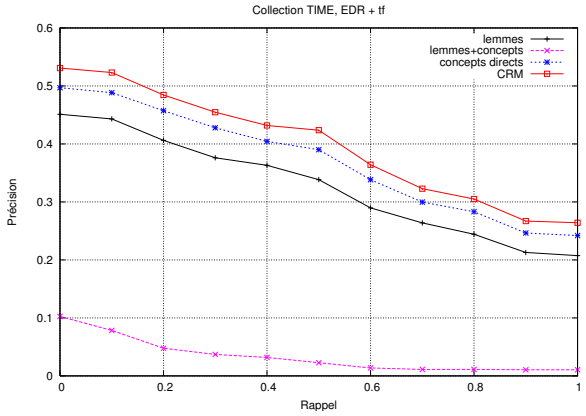
[1] Part-of-Speech tagging, which will be studied in more details later in the semester, consists in adding each word a ("Part-of-Speech") tag corresponding to its syntactic role within the sentence:.

⑥ Another very general evaluation framework concerns this kind of NLP tasks where the goal of the system is to propose a set of outputs among which some might turn to be correct, while other might not (e.g. Information Retrieval (IR)). In this type of situation, the standard evaluation metrics are the Precision and the Recall.

Give the formal definition of Precision and Recall and indicate some examples of NLP tasks (other than IR) that can be evaluated with the Precision/Recall metrics.

⑦ Consider the following Precision/Recall curves





What conclusions can one derive from such curves? Provide a detailed interpretation of the results.

⑧ It is often desirable to be able to express the performance of an NLP system in the form of one single number, which is not the case with Precision/Recall curves.

Indicate what score can be used to convert a Precision/Recall performance into a unique number. Give the formula for the corresponding evaluation metric, and indicate how it can be weighted.

⑨ Give well chosen examples of applications that can be evaluated with the single metric derived from Precision/Recall and illustrate:

- a situation where more weight should be given to Precision;
- a situation where more weight should be given to Recall.

Exercise II.2

You have been hired to *evaluate* an email monitoring system aimed at detecting potential security issues. The targeted goal of the application is to decide whether a given email should be further reviewed or not.

① Give four standard measures usually considered for the evaluation of such a system? Explain their meaning. Briefly discuss their advantages/drawbacks.

② For three of the measures you mentioned in the previous question, what are the corresponding scores for a system providing the following results:

email	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}
referential	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_2	C_2
system	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_1	C_2	C_1	C_2	C_1	C_2	C_2

③ You have been given the results of three different systems that have been evaluated on the same panel of 157 different emails. Here are the classification errors and their standard deviations:

	system 1	system 2	system 3
error	0.079	0.081	0.118
std dev	0.026	0.005	0.004

Which system would you recommend? Why?

④ **Optional** (too advanced for the current version of the course)**:** What should be the minimal size of a test set to ensure, at a 95% confidence level, that a system has an error 0.02 lower (absolute difference) than system 3? Justify your answer.

4 Tokenization/Lexicons/n-grams

Exercise IV.1

According to your knowledge of English, split the following sentence into words and punctuation:

M. O'Connel payed $ 12,000 (V.T.A. not included) with his credit card.

Which of these words won't usually be in a standard lexicon? Justify your answer.

Assuming separators are: whitespace, quote ('), full-stop/period (.), parenthesis,
and that separators a kept as tokens, tokenize the former sentence.

How would you propose to go from tokens to words? (propose concreat implementations)

Exercise IV.2

Consider the following toy corpus:

the cat cut the hat

- How many different bigrams of characters (including whitespace) do you have in that corpus?
- How many occurences do you have in total? (i.e. including repertitions)
- Considering only lowercase alphabetical and whitespace, how many bigrams are possible?
- What are the parameters of a bigram model using the same set of characters (lowercase alphabetical and whitespace)?
- What is the probability of the following sequences, if the parameters are estimated using MLE (maximum-likelihood estimation) on the above corpus (make use of a calculator or even a short program):
  - *cutthechat*
  - *cut the chat*

  Fully justify your answer.
- What is the probability of the same sequences, if the parameters are estimated using Dirichlet prior with α having all its components equal to 0.05?

  Fully justify your answer.

5 Part-of-Speech tagging

Exercise V.1

What is the tagging of the following sentence

computers process programs accurately

with the following HMM tagger:

(part of) lexicon:

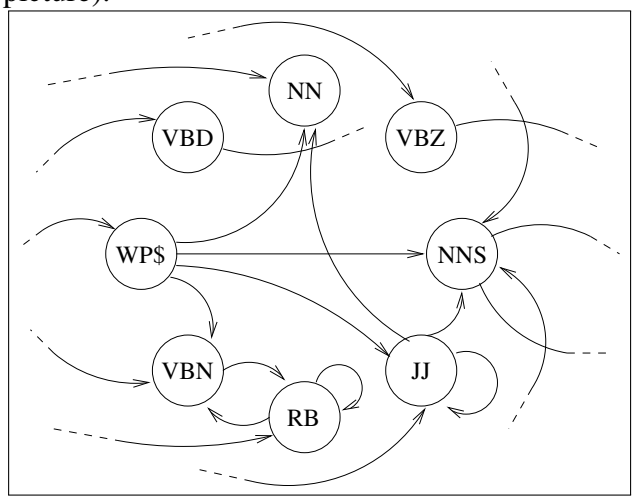
```
computers   N       0.123
process     N       0.1
process     V       0.2
programs    N       0.11
programs    V       0.15
accurately  Adv     0.789
```

(part of) transitions:

```
P(N|V)=0.5      P(N|Adv)=0.12       P(V|Adv)=0.05
P(V|N)=0.4      P(Adv|N)=0.01       P(Adv|V)=0.13
P(N|N)=0.6      P(V|V)=0.05
```

Exercise V.2

We aim at tagging English texts with "Part-of-Speech" (PoS) tags. For this, we consider using the following model (partial picture):



continues on back ☞

Explanation of (some) tags:

Tag	English expl.	Expl. française	Example(s)
JJ	Adjective	adjectif	*yellow*
NN	Noun, Singular	nom commun singulier	*cat*
NNS	Noun, Plural	nom commun pluriel	*cats*
PRP$	Possessive Pronoun	pronom possessif	*my*, *one's*
RB	Adverb	adverbe	*never*, *quickly*
VBD	Verb, Past Tense	verbe au passé	*ate*
VBN	Verb, Past Participle	participe passé	*eaten*
VBZ	Verb, Present 3P Sing	verbe au présent, 3e pers. sing.	*eats*
WP$	Possessive wh-	pronom relatif (poss.)	*whose*

① What kind of model (of PoS tagger) is it? What assumption(s) does it rely on?

② What are its parameters? Give examples and the appropriate name for each.

We use the following (part of) lexicon:

adult	JJ	has	VBZ
adult	NN	just	RB
daughter	NN	my	PRP$
developed	VBD	programs	NNS
developed	VBN	programs	VBZ
first	JJ	tooth	NN
first	RB	whose	WP$

and consider the following sentence:

my daughter whose first adult tooth has just developed programs

③ With this lexicon, how many different PoS taggings does this sentence have? Justify your answer.

④ What (formal) parameters make the difference in the choice of these different PoS taggings (for the above model)?

Give the explicit mathematical formulas of these parts that are different.

⑤ Assume that the following tagging is produced:

my/PRP$ daughter/NN whose/WP$ first/JJ adult/JJ tooth/NN has/VBZ just/RB developed/VBN programs/NNS

How is it possible? Give an explanation using the former formulas.

Exercise V.3

① What is the problem addressed by a Part-of-Speech (PoS) tagger?
Why isn't it trivial? What are the two main difficulties?

② Assume that the texts to be tagged contain unknown words, which are either capitalized words, or spelling errors, or simply general common words not seen during the learning. Almost all capitalized words correspond to proper nouns, and most of the spelling-errors correspond to words already in the lexicon (only a few of the spelling errors correspond to words not seen during the learning).

How would you handle such a situation in a concrete NLP application (that uses a PoS tagger)? Explicit your solution(s).

③ Assume that the texts to be tagged contain 1.5% of unknown words and that the performance of the tagger to be used is 98% on known words.

What will be its typical overall performance in the following two situations:

(a) all unknown words are systematically wrongly tagged?

(b) using the solution you proposed in ② is used in a situation where 80% of the unknown words are capitalized among which 98% are proper nouns, 15% are general common words not seen during learning, and 5% are spelling-errors, among which 1% corresponds to correct words which were not in the learning set?

Provide both a calculation (a complete formula but not necessarily the final numerical result) and an explanation.

Exercise V.4

① Consider an HMM Part-of-Speech tagger, the tagset of which contains, among others:

$$\texttt{DET}, \texttt{N}, \texttt{V}, \texttt{ADV} \text{ and } \texttt{ADJ},$$

and some of the parameters of which are:

$$\begin{array}{lll} P_1(\text{a}|\texttt{DET}) = 0.1, & P_1(\text{accurately}|\texttt{ADV}) = 0.1, & P_1(\text{computer}|\texttt{N}) = 0.1, \\ P_1(\text{process}|\texttt{N}) = 0.095, & P_1(\text{process}|\texttt{V}) = 0.005, & \\ P_1(\text{programs}|\texttt{N}) = 0.080, & P_1(\text{programs}|\texttt{V}) = 0.020, & \end{array}$$

continues on back ☞

$P_2(\texttt{Y}|\texttt{X})$: (for instance $P_2(\texttt{N}|\texttt{DET}) = 0.55$)

		Y →				
		DET	N	V	ADJ	ADV
X ↓	DET	0	0.55	0	0.02	0.03
	N	0.01	0.10	0.08	0.01	0.02
	V	0.16	0.11	0.06	0.08	0.08
	ADJ	0.01	0.65	0	0.05	0
	ADV	0.08	0.02	0.09	0.04	0.04

and:
$P_3(\texttt{DET}) = 0.20$, $P_3(\texttt{N}) = 0.06$, $P_3(\texttt{V}) = 0.08$, $P_3(\texttt{ADV}) = 0.07$, $P_3(\texttt{ADJ}) = 0.02$.

(a) How are the propabilities P_1, P_2 and P_3 usually called?

P_1:

P_2:

P_3:

(b) What are all the possible taggings of the sentence

`a computer process programs accurately`

(c) What would be the output of the HMM PoS tagger on the above sentence? Fully justify your answer.

7 Text Classification

Exercise VII.1

In an automated email router of a company, we want to make the distinction between three kind of emails: technical (about computers), financial, and the rest ("irrelevant"). For this we plan to use a Naive Bayes approach.

① What is the main assumption made by Naive Bayes classifiers? Why is it "Naive"?

We will consider the following three messages:

> The Dow industrials tumbled 120.54 to 10924.74, hurt by GM's sales forecast and two economic reports. Oil rose to $71.92.

from `www.wsj.com/`

> BitTorrent Inc. is boosting its network capacity as it prepares to become a centralized hub for legal video content. In May, BitTorrent announced a deal with Warner Brothers to distribute its TV and movie content via the BT platform. It has now lined up IP transit for streaming videos at a few gigabits per second

from `slashdot.org/`

> Intel will sell its XScale PXAxxx applications processor and 3G baseband processor businesses to Marvell for $600 million, plus existing liabilities. The deal could make Marvell the top supplier of 3G and later smartphone processors, and enable Intel to focus on its core x86 and wireless LAN chipset businesses, the companies say.

from `www.linuxdevices.com/`

② What pre-processing steps (before actually using the Naive Bayes Classifier) do you consider applying to the input text?

③ For the first text, give an example of the corresponding output of the pre-processor.

continues on back ☞

Suppose we have collected the following statistics[2] about the word frequencies within the corresponding classes, where "0.00…" stands for some very small value:

	technical	financial	irrelevant
$<number>	0.01	0.07	0.05
Dow	0.00…	0.08	0.00…
GM	0.00…	0.03	0.00…
IP	0.03	0.00…	0.00…
Intel	0.02	0.02	0.00…
business	0.01	0.07	0.04
capacity	0.01	0.00…	0.00…
chipset	0.04	0.01	0.00…
company	0.01	0.04	0.05

	technical	financial	irrelevant
deal	0.01	0.02	0.00…
forecast	0.00…	0.03	0.01
gigabit	0.03	0.00…	0.00…
hub	0.06	0.00…	0.01
network	0.04	0.01	0.00…
processor	0.07	0.01	0.00…
smartphone	0.04	0.04	0.01
wireless	0.02	0.01	0.00…

④ For each of the above three texts, in what category will it be classified, knowing that on average 50% of the emails happen to be technical, 40% to be financial and 10% to be of no interest.

You can assume that all the missing information is irrelevant (i.e. do not impact the results).

Provide a full explanation of all the steps and computations that lead to your results.

We now want to specifically focus on the processing of compounds such as "network capacity" in the second text.

⑥ How are the compounds handled by a Naive Bayes classifier if no specific pre-processing of compounds is used?

⑦ What changes if the compounds are handled by the NL pre-processor?

Discuss this situation (NL pre-processing handling compounds) with respect to the Naive Bayes main assumption.

⑧ Outline how you would build a pre-processor for compound words.

[2]Note that this is only partial information, statistics about other words not presented here have also been collected.

Exercise VII.2

You are responsible for a project aiming at providing on-line recommendations to the customers of a on-line book selling company.

The general idea behind this recommendation system is to cluster books according to both customers and content similarities, so as to propose books similar to the books already bought by a given customer. The core of the recommendation system is a clustering algorithm aiming at regrouping books likely to be appreciate by the same person. This clustering should not only be achieved based on the purchase history of customers, but should also be refined by the content of the books themselves. It's that latter aspect we want to address in this exam question.

① Briefly explain how books could be clustered according to similar content. Give the main steps and ideas.

② The chosen clustering algorithm is the dendrogram. What other algorithms could you propose for the same task? Briefly review advantages and disadvantages of each of them (including dendrograms). Which one would you recommend for the targeted task?

③ Consider the following six "documents" (toy example):

d_1 "*Because cows are not sorted as they return from the fields to their home pen, cow flows are improved.*"

d_2 "*He was convinced that if he owned the fountain pen that he'd seen in the shop window for years, he could write fantastic stories with it. That was the kind of pen you cannot forget.*"

d_3 "*With this book you will learn how to draw humans, animals (cows, horses, etc.) and flowers with a charcoal pen.*"

d_4 "*The cows were kept in pens behind the farm, hidden from the road. That was the typical kind of pen made for cows.*"

d_5 "*If Dracula wrote with a fountain pen, this would be the kind of pen he would write with, filled with blood red ink. It was the pen she chose for my punishment, the pen of my torment. What a mean cow!*"

d_6 "*What pen for what cow? A red pen for a red cow, a black pen for a black cow, a brown pen for a brown cow, ... Understand?*"

and suppose (toy example) that they are indexed only by the two words: *pen* and *cow*.

(a) Draw their vector representations.

(b) Give the definition of the cosine similarity. What vector's feature(s) is it sensible to?

(c) What is the result of the dendrogram clustering algorithm on those six documents, using the cosine similarity and single linkage?
Explain all the steps.

Hint: $5/\sqrt{34} < 3/\sqrt{10} < 4/\sqrt{17}$.

8 Information Retrieval

Exercise VIII.1

① Describe the main principles of the standard vector space model for semantics.

② Consider the following document:

D = "`the exports from Switzerland to the USA are increasing in 2006`"

Propose a possible indexing set for this document. Justify your answer.

③ What is the similarity between the above document D and

D' = "`Swiss exports have increase this year`"

Justify your answer.

④ Briefly describe the important limitation(s) of the standard vector space approach.

Explain how more sophisticated techniques such as the *Distributional Semantics* can be used to circumvent this/these limitation(s).

⑤ Give some concrete examples of NLP applications that might benefit from the semantic vectorial representations.

⑥ Using the standard vector space model, does the indexing set you considered in question Q2 allow to discriminate between D and this other document:

D" = "`the exports from the USA to Switzerland are increasing in 2006`"

If yes: how? If not, why?

⑦ Would a parser be available, how could it be used to provide a (partial) solution to the problem?

Exercise VIII.2

① What is the cosine similarity?

② Consider the following two documents:

```
D1: Dog eat dog. Eat cat too!
D2: Eat home, it's raining cats and dogs.
```

What would be their cosine similarity in a typical information retrieval setup? Explain all the steps.

③ In a standard cosine-based tf-idf information retrieval system, can a query retrieve a document that does not contain any of the query words? Justify your answer.

④ Other measures than the cosine are possible. For instance, Jaccard similarity computes the ratio between the number of words in common and the total number of occurring words (i.e. "intersection over union").

(a) Can this measure be used on the same document representations as the one used for the cosine similarity can be? Justify your answer.

(b) Using the boolean representation for documents, find an example of three distinct documents d_1, d_2 and d_3 such that d_1 is closest to d_2 using cosine similarity, whereas d_2 and d_3 have the same similarity with respect to d_1 using Jaccard similarity.

Conclude on the comparison between these two similarity measures.

⑤ An information retrieval system with high precision and low recall can be useful for:

(a) Retrieving all relevant documents from a database of legal cases.

(b) Retrieving some interesting documents for a given a topic from the web.

(c) Retrieving a large set of interesting documents for a given a topic from the web.

(d) Checking the existence of a document in a very large document collection.

Choose all possible useful situations in the above list (maybe several). Justify your answer; in particular, define the notions of precision and recall.

Exercise VIII.3

① Official NLP evaluations (especially for task such as Information Retrieval or Information Extraction) are often carried out in the form of "evaluation campaigns".

Precisely describe the various steps of such an evaluation campaign.

For each of the steps, clearly indicate the main goals.

② In an IR evaluation campaign, the following "referential" ("golden truth") has been produced by a set of human judges:

```
q1: d01 d02 d03 d04
q2: d05 d06
q3: d07 d08 d09 d10 d11
q4: d12 d13 d14 d15
```

where the list of document references `dj` associated with a query reference `qi` defines the set of documents considered to be relevant for the query by the human judges.

Is such a referential easy to produce?

Indicate the various problems that might arise when one tries to produce it.

③ Consider two Information Retrieval systems S_1 and S_2 that produced the following outputs for the 4 reference queries q1, q2, q3, q4:

```
S1:                                          | referential:
 q1: d01 d02 d03 d04 dXX dXX dXX dXX         |  q1: d01 d02 d03 d04
 q2: d06 dXX dXX dXX dXX                     |  q2: d05 d06
 q3: dXX d07 d09 d11 dXX dXX dXX dXX dXX     |  q3: d07 d08 d09 d10 d11
 q4: d12 dXX dXX d14 d15 dXX dXX dXX dXX     |  q4: d12 d13 d14 d15

S2::                                         | referential:
 q1: dXX dXX dXX dXX d04                     |  q1: d01 d02 d03 d04
 q2: dXX dXX d05 d06                         |  q2: d05 d06
 q3: dXX dXX d07 d08 d09                     |  q3: d07 d08 d09 d10 d11
 q4: dXX d13 dXX d15                         |  q4: d12 d13 d14 d15
```

where dXX refer to document references that do not appear in the referential. To make the answer easier, we copied the referential on the right.

For each of the two systems, compute the mean Precision and Recall measures (provide the results as fractions). Explain all the steps of your computation.

④ Explain how it is possible to compute Precision at different Recalls.

⑤ How is it possible to compute the average Precision/Recall curves? Explain in detail the various steps of the computation.

As it would be too tedious to compute the average Precision/Recall curves by hand, plot, on a Precision/Recall graph, the Precision and Recall values obtained in subquestion ③ for each of the two systems and for each of the 4 queries.

Based on the resulting curves, what is your relative evaluation of the two systems?

⑥ The Precision/Recall based evaluation of the IR systems S1 and S2 above does not explicitly take into account the order in which the documents have been retrieved by the systems. For this purpose, another metric can be used: the Precision at k (P@k), which corresponds to the fraction of truly relevant documents among the top k documents retrieved by a system.

Compute the average P@k values for k between 1 and 5 for the IR systems S1 and S2 above. What additional insight do these values provide in addition to the Precision/Recall curves?

Based on these results, what is your relative evaluation of the two systems? How does it compare to ③?

⑦ It is often desirable to be able to express the performance of an NLP system in the form of a single number, which is not the case when the Precision/Recall framework is used.

Indicate what scores can be used to convert Precision/Recall measures into a unique number. For each score, give the corresponding formula.

⑧ Give well chosen examples of applications that illustrate:

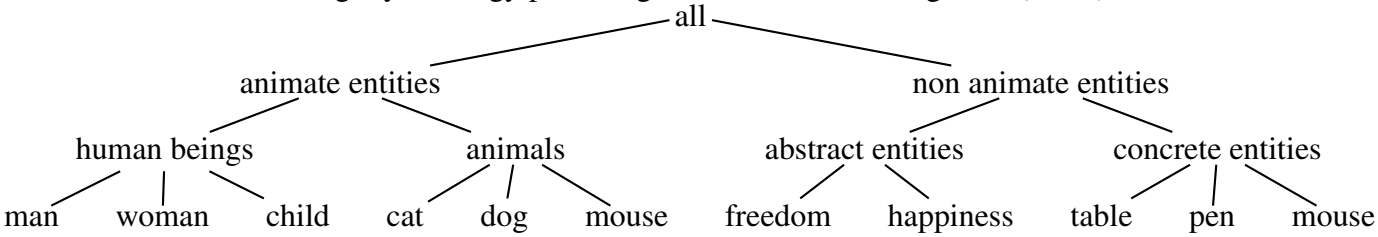
- a situation where more importance should be given to Precision;
- a situation where more importance should be given to Recall.

9 Lexical Semantics

Exercise IX.1

The objective of this question is to illustrate the use of a lexical semantics resource to compute lexical cohesion.

Consider the following toy ontology providing a semantic structuring for a (small) set of nouns:



① What is the semantic relation that has been used to build the ontology?

Cite another semantic relations that could also be useful for building lexical semantics resources.

For this semantic relation, give a short definition and a concrete example.

② The word "*mouse*" appears at two different places in the toy ontology. What does this mean?

What specific problems does it raise when the ontology is used?

How could such problems be solved? (just provide a sketch of explanation.)

③ Consider the following short text:

```
Cats are fighting dogs. There are plenty of pens on the table.
```

What pre-processing should be performed on this text to make it suitable for the use of the available ontology?

④ We want to use lexical cohesion to decide whether the provided text consists of one single topical segment corresponding to both sentences, or of two distinct topical segments, each corresponding to one of the sentences.

Let's define the lexical cohesion of any set of words (in canonical form) as the average lexical distance between all pairs of words present in the set[3]. The lexical distance between any two words is be defined as the length of a shortest path between the two words in the available ontology.

For example, "*freedom*" and "*happiness*" are at distance 2 (length, i.e. number of links, of the path: happiness $\longrightarrow$ abstract entities $\longrightarrow$ freedom), while "*freedom*" and "*dog*" are at distance 6 (length of the path: freedom $\longrightarrow$ abstract entities $\longrightarrow$ non animate entities $\longrightarrow$ all $\longrightarrow$ animate entities $\longrightarrow$ animals $\longrightarrow$ dog)

Compute the lexical distance between all the pairs of words present in the above text and in the provided ontology (there are 6 such pairs).

[3]Here, is actually the *lack of* cohesion that we measure: since it's a distance, the lower the more cohesion and the bigger the less cohesion.

⑤ Compute the lexical cohesion of each of the two sentences, and then the lexical cohesion of the whole text.

Based on the obtained values, what decision should be taken as far as the segmentation of the text into topical segments is concerned?

⑥ Give some examples of NLP tasks for which lexical cohesion might be useful. Explain why.