Lecture reviews — Week 03
with solutions

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Purpose of these lecture reviews

- Improve/deepend your learning
- Answer your questions
- Save you practice/revision time

Why are these sessions not recorded?

1. the intention is to have *appropriate*/*adapted*/*personalized* face-to-face interaction
2. recording them would lead to an extra 2 hours/week video lecture
   (which is too much *passive* content)

Content

1. Big picture:
   What did you retain? What keypoints do you remember?
2. Questions?
3. More examples

Week 3 keypoints

- Words vs. tokens
- Role of a lexicon
- Storage of surface form field
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)

Questions?

Week 3 practice example 1

What is the expected output of a *standard* tokenizer when applied to the following title:

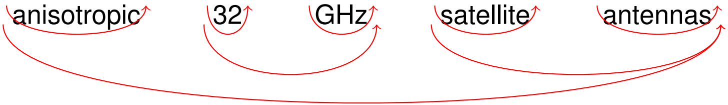
anisotropic | 32 | GHz | satellite | antennas | for | high-speed | 5G | networks

What does “standard” mean anyway? Is the hyphen a standard separator?
☞ define at least your separators

To improve the quality of the tokenization, you decide to use a lexicon containing all possible word forms occurring in the titles, including compounds, such as “*32 GHz*”, “*satellite antennas*”, or “*5G networks*”.

- What is a possible approach allowing to efficiently implement such a lexicon, if we want to guarantee a constant time access to the entries of the lexicon, and the possibility to use regular expressions as lexical entries? ☞ FSA
- Indicate the result of all *usefull* tokenizations of the following string by drawing all the generated arcs on top of that string:

What does “usefull” mean anyway? ☞ depends on the application


anisotropic 32 GHz satellite antennas

Week 3 practice example 2

Take a random Wikipedia page (e.g. `https://en.wikipedia.org/wiki/ACVRL1`) and compare two phrases using 3-grams (of tokens).
For instance:
This gene encodes a type I receptor
and
This gene encodes a type 2 receptor

1. Where to start from (in the corpus/in the document)?
   ☞ meta-information do help!
2. What words/tokens? (e.g. "*Serine/threonine-protein kinase recept*")
   Pay also attention to meaningful specificities, e.g. what about "type II receptor"?
3. How to deal with upper-/lowercase? (e.g. "*This*")
   Notice that $P(\text{This})$ is in fact $P(\text{this}|<\texttt{BoS}>)$
4. What estimates? (MLE? Smoothing?)