Lecture reviews — Week 04 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 4 keypoints

- Words vs. tokens
- *n*-gram models
- MLE and add-one smoothing are bad (in NLP)
- Language Identification
- Out-of-Vocabulary froms:
  - OoV forms do matter
  - 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors

Questions?

Week 4 review example

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?) Smoothing, for sure! For instance:

$$P(n-\text{gram}) = \frac{\text{count}+\alpha}{N+M\alpha}$$

N: number of occurences in the learning corpus (typically: size of corpus - $n+1$)
M: number of possible n-grams (typically some m^n)

Week 4 review example – Hints

- What do we want to do first?
  ☞ estimate a 3-gram language model (of tokens)

- What is the first parameter estimated?
  Assuming we answered the first three points of the former slide by (this is *just* one possible choice):
  1. consider only “main full text” (ignore all other infos)
  2. tokenize on [A-Za-z0-9] only
  3. lowercase + sentence detection (<BoS>)

  then, the first estimated parameter will be: $P(<\mathtt{BoS}>, serine, /)$

- Finaly use parameters to compare the two sequences.
  In this very case, this ends up to comparing
  $P(\text{I}|\text{a type}) \cdot P(\text{receptor}|\text{type I})$
  with
  $P(\text{2}|\text{a type}) \cdot P(\text{receptor}|\text{type 2})$