

Lecture reviews — Week 04

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Purpose of these lecture reviews

- ▶ Improve/deepen 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 (deepened together)

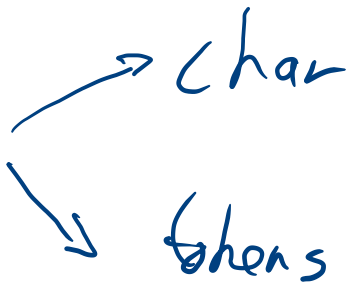
Week 4 keypoints

- tokenization : split

token / word



- n-grams



parameter

estimated : MLE /

Smoothing → add
↓
Dirichlet prior

- OoV

Week 4 keypoints

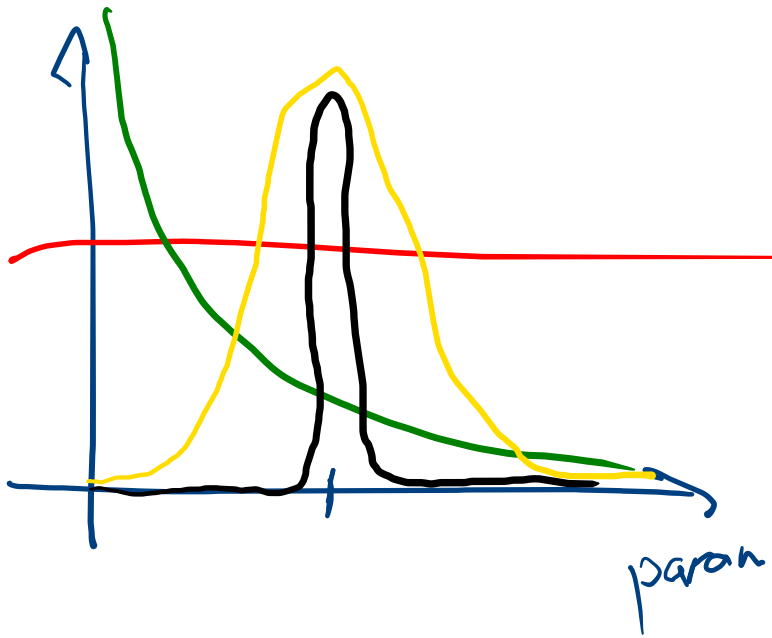
- ▶ Words vs. tokens
- ▶ n -gram models
- ▶ MLE and add-one smoothing are bad (in NLP)
- ▶ Language Identification
- ▶ Out-of-Vocabulary forms:
 - ▶ OoV forms do matter
 - ▶ 4 types of OoV: neologisms, borrowings, forms difficult to lexicalize, spelling errors

$$P(\overset{\text{para}}{\cancel{x}} | Y) = \frac{P(Y | \overset{\text{para}}{\cancel{x}}) \cdot P(\overset{\text{para}}{\cancel{x}})}{P(Y)}$$

$$\text{para.} : P(y_1 y_2 y_3)$$

$$\tilde{P}(y) = \frac{\#y + \alpha_i}{N + \sum_k \alpha_k}$$

$$P(\text{para}) \propto \text{Dirichlet}(\alpha_1 \dots \alpha_H)$$



$$\alpha_1 = \alpha_2 = \dots = \alpha_k = 1$$

Week 4 keypoints

- ▶ Words vs. tokens
- ▶ n -gram models
- ▶ MLE and add-one smoothing are bad (in NLP)
- ▶ Language Identification
- ▶ Out-of-Vocabulary forms:
 - ▶ 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:

$$P(\text{This gene encodes a type 1 receptor}) = P(\text{This gene encodes}) \cdot P(a | \text{gene encodes})$$

and

$$P(\text{This gene encodes a type 2 receptor}) = \dots \cdot P(\text{receptor type 1})$$

Handwritten note: param? P (the cat mouse)

This

→ this | <Bos>

estimation?

$$\tilde{P}(\text{This gene encodes}) = \frac{\# \text{ times "This gene encodes" appears}}{\# \text{ trigrams (of tokens)}}$$

$$\alpha_i = 0.5$$

$$= \frac{1 + 0.5}{\# \text{ token} - 2 + T^3 \cdot 0.5}$$

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 1 receptor

and

This gene encodes a type 2 receptor

1. Where to start from (in the corpus/in the document)?

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 1 receptor

and

This gene encodes a type 2 receptor

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)

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 1 receptor

and

This gene encodes a type 2 receptor

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)
3. How to deal with upper-/lowercase? (e.g. “*This*”)

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 1 receptor

and

This gene encodes a type 2 receptor

1. Where to start from (in the corpus/in the document)?
2. What words/tokens? (e.g. “*Serine/threonine-protein kinase recept*”)
3. How to deal with upper-/lowercase? (e.g. “*This*”)
4. What estimates? (MLE? Smoothing?)