Lecture reviews — Week 05

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

Week 5 keypoints

$$P(w_i \mid \text{whatever containing } t_i) = P(w_i \mid t_i)$$

lexical

$$P(t_i \mid t_1 \ldots t_{i-1}) = P(t_i \mid t_{i-1})$$

syntactic

$$\sum_t P(t \mid t') = 1$$

Order 2

Week 5 keypoints

- what "lemmatization" is
  - some kind of normalisation of the surface-forms
  - lematization is made easier once PoS-tagging has been done
  - otherwise: "stemmer"
- what "part-of-speech tagging" is

  to choose the right tag *according to the context*, among the possible PoS-tags for each word of the input text
- two hypothesis to transform PoS tagging into "the second problem" of HMMs
  - limited lexical conditioning:
    $P(w_i|w_1, ..., w_{i-1}, t_1, ..., t_i, ..., t_n) = P(w_i|t_i)$
  - k-neighbors limited scope for syntactic dependencies:
    $P(t_i|t_1, ..., t_{i-1}) = P(t_i|t_{i-k}, ..., t_{i-1})$
- order of magnitude of performances
  95–99% (random: 75–90%)

Week 5 practice example (1/2)

initial   transition   emit

① Consider an order-1 HMM PoS tagger using a lexicon with N entries, and a tag set with T tags. Furthermore, assume that the entries of the lexicon are associated, on the average, with t distinct tags.
Provide (an estimate of) the total number Q of (not necessarily free) parameters to be estimated to exploit the order-1 HMM model, assuming that no guesser has been implemented. **Justify** your answer.

$T \quad T^2 \quad N \cdot \cancel{T}\, t$

$Q \simeq N(t+1)$

② Consider the following lexicon excerpt, where D, N, P, and V are the tags associated with the entries
(D stands for determiner, N for noun, P for pronoun, and V for verb):

cat: N, V	saw: N, V
run: N, V	the: D
running: N, V	you: P

Provide and justify the number M of potential PoS taggings that have to be considered for the following sentence:

the cat you saw running

D N P V V

$M = 1 \times 2 \times 1 \times 2 \times 2$

J.-C. Chappelier & M. Rajman

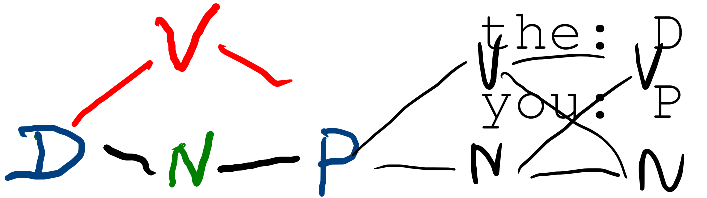
Week 5 practice example (2/2)

emit $P(cat|N)$

init $P(D)$

trans $P(N|D)$

cat:	N, V
run:	N, V
running:	N, V
saw:	N, V
the:	D
you:	P



the cat you saw running

③ What is the condition to be verified by the parameters of the order-1 HMM model (using the provided lexicon excerpt) for the word "cat" to be tagged as a noun in the above sentence?
Justify your answer.

$$\cancel{P(the|D)}\ \cancel{P(D)}\ P(N|D)\cdot P(cat|N)\cdot P(P|N) >$$

$$\cancel{P(the|D)}\ \cancel{P(D)}\ P(V|D)\cdot P(cat|V)\cdot P(P|V)$$

$$\max_{t_1^n} P(t_1) \cdot P(w_1|t_1) \cdot \prod_{i=2}^{n} P(t_i|t_{i-1}) \cdot P(w_i|t_i)$$

$$\underbrace{\qquad\qquad\qquad}_{A} \quad \Big| \quad P(w_k|t_k) \quad \underbrace{\qquad\qquad}_{B}$$