Lecture reviews — Week 05

J.-C. Chappelier & M. Rajman

Laboratoire d'Intelligence Artificielle
Faculté I&C

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%)

$$P(t_1 \dots t_n \mid \omega_1 \dots \omega_n)$$

lexical

$$= \frac{P(\omega_1^n \mid t_1^n) \cdot P(t_1^n)}{P(\omega_1^n)}$$

syntax

$$\prod_i P(\omega_i \mid t_i)$$

Week 5 practice example (1/2)

P(tag to start) : P(word/tag) : P(tag/prec.tag)
init : emit : transit

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 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.

$Q =$ #init + #emit + #transition $= T + (\simeq t \cdot N) + T^2$

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
run: N, V
running: N, V
saw: N, V
the: D
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
```

$M =$ 1 2 1 2 2 $= 2^3 = 8$

Week 5 practice example (2/2)

the cat you saw running

D N P N N

V V

non ambiguous

```
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.

$P(D\ N\ P\ \ldots \mid \text{the cat you saw running})$

$= P_{init}(D) \cdot P(\text{the} \mid D) \cdot P(N \mid D) \cdot P(\text{cat} \mid N) \cdot P(P \mid N) \cdot P(\text{you} \mid P) \ldots$

emit transition the rest

alternative $P(DVP\ldots) = P(D) \cdot P(the|D) \cdot P(V|D) \cdot P(cat|V) \cdot P(P|V) \cdot P(you|P)$

$P(DVP\ldots) < P(DNP\ldots)$

the same rest

$P(V|D) \cdot P(cat|V) \cdot P(P|V) < P(N|D) \cdot P(cat|N) \cdot P(P|N)$

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

Faculté Informatique et Communication
Introduction to Natural Language Processing (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Part-of-Speech tagging (part 1)

J.-C. Chappelier M. Rajman

v. 20211020 – 1

QUESTION I [1 pt]

(from Fall 2018 quiz 2)

Assume that, in the word sequence "*iron shaped cloth*", the word "*shaped*" is replaced by an Out-of-Vocabulary (OoV) form that your spell checker was not able to correct, nor your morphological analyzer to analyze. Select among the following options the most adequate one to decide which possible PoS tags should be associated with the OoV form:

[ ] All the PoS tags

[ ] All the PoS tags corresponding to open grammatical categories

[ ] The most frequent PoS tag

QUESTION II [2 pt]

(from Fall 2018 quiz 2)

For this question, *one or more* assertions can be correct. Tick only the correct assertion(s). There will be a penalty for wrong assertions ticked.

When using a probabilistic approach to find the optimal tagging for the sentence "*young birds fly*", what does the conditional probability

$$P(\text{“}birds\text{”}, \text{“}fly\text{”} \mid t_2\, t_3)$$

represent, provided that no additional specific hypotheses are made:

- [ ] the probability that the word *“birds”* appears at position 2 conditioned by the tag t_2 only, and the word *“fly”* appears at position 3 conditioned by tag t_3 only;
- [ ] the probability that the word *“birds”* and the tag t_2 appear at position 2, and the word *“fly”* and the tag t_3 appear at position 3;
- [ ] the probability that the sequence (*“birds”*, *“fly”*) appears at the end of the sentence, conditioned by the tag pair (t_2 , t_3).

QUESTION III [2 pt]

(from Fall 2018 quiz 2)

For this question, *one or more* assertions can be correct. Tick only the correct assertion(s). There will be a penalty for wrong assertions ticked.

Indicate which of the following formulas are strictly equal to the conditional probability

$$P(\text{“}young\text{”}, \text{“}birds\text{”}, \text{“}fly\text{”} \mid t_1, t_2, t_3)$$

provided that no specific additional hypotheses are made:

- [ ] $P(\text{“}young\text{”} \mid t_1) \cdot P(\text{“}birds\text{”} \mid t_2, \text{“}young\text{”}, t_1) \cdot P(\text{“}fly\text{”} \mid t_3, \text{“}young\text{”}, t_1, \text{“}birds\text{”}, t_2)$
- [ ] $P(\text{“}young\text{”} \mid t_1, t_2, t_3) \cdot P(\text{“}birds\text{”} \mid \text{“}young\text{”}, t_1, t_2, t_3) \cdot P(\text{“}fly\text{”} \mid \text{“}young\text{”}, \text{“}birds\text{”}, t_1, t_2, t_3)$
- [ ] $P(\text{“}young\text{”} \mid t_1) \cdot P(\text{“}birds\text{”} \mid t_2) \cdot P(\text{“}fly\text{”} \mid t_3)$
- [ ] $P(\text{“}young\text{”} \mid t_1) \cdot P(\text{“}birds\text{”} \mid t_2) \cdot P(\text{“}fly\text{”} \mid t_3) \cdot P(t_1) \cdot P(t_2 \mid t_1) \cdot P(t_3 \mid t_2)$

QUESTION IV [2 pt]

(from Fall 2018 quiz 2)

For this question, *one or more* assertions can be correct. Tick only the correct assertion(s). There will be a penalty for wrong assertions ticked.

When using an HMM to find the optimal tagging of a 3 word sentence "$w_1\ w_2\ w_3$", what probability should be maximized?

[ ] $P(w_1, w_2, w_3|t_1, t_2, t_3)$

[ ] $P(t_1, t_2, t_3|w_1, w_2, w_3)$

[ ] $P(w_1, w_2, w_3|t_1, t_2, t_3) \cdot P(t_1, t_2, t_3)$

[ ] $P(w_1|t_1, t_2, t_3) \cdot P(w_2|w_1, t_1, t_2, t_3) \cdot P(w_3|w_1, w_2, t_1, t_2, t_3)$

QUESTION V [2 pt]

(from Fall 2018 quiz 2)

① Cross out the elements that can be ignored in the following conditional probability, when tagging the sentence "*young birds fly*" under the "*limited lexical conditioning*" hypothesis:

$$P(\text{“}fly\text{”} \mid \text{“}young\text{”}, \text{“}birds\text{”}, t_1, t_2, t_3)$$

② Cross out the elements that can be ignored in the following conditional probability, when tagging the sentence "*young birds fly fast*" under the "*limited scope for syntactic dependencies (1 neighbor)*" hypothesis:

$$P(t_3 \mid t_1, t_2, t_4)$$

QUESTION VI [4 pt]

(from Fall 2018 quiz 2)

For this question, *one or more* assertions can be correct. Tick only the correct assertion(s). There will be a penalty for wrong assertions ticked.

① When using an HMM to find the optimal tagging of the sentence "*young ducks fly fast*", where all the words are considered as potentially ambiguous, indicate which of the following assertions are true under the "*limited lexical conditioning*" and "*limited scope for syntactic dependencies (1 neighbor)*" hypotheses

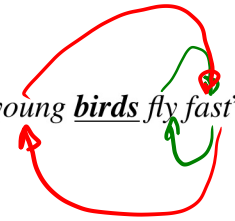
[ ] The tagging of "*fast*" depends on the one of "*fly*".

[ ] The tagging of "*fly*" depends on the one of "*fast*".

[ ] The tagging of "*young*" depends on the one of "*fast*".

[ ] The tagging of "*fast*" depends on the one of "*young*".

② Same question, but for the sentence "*young* ***birds*** *fly fast*", where all the words but "*birds*" are considered as potentially ambiguous:



[ ] The tagging of "*fast*" depends on the one of "*fly*".

[ ] The tagging of "*fly*" depends on the one of "*fast*".

[ ] The tagging of "*young*" depends on the one of "*fast*".

[ ] The tagging of "*fast*" depends on the one of "*young*".