IHLT Exam Q2-2018/2019
Start time 10:00; End time 13:00

1. (2 points) Given the following sequences of pairs (state, emission):

(D,the) (N,wine) (V,ages) (A,alone) (FF, .)
(D,the) (N,wine) (N,waits) (V,last) (N,ages) (FF, .)
(D,some) (N,flies) (V,dove) (P,into) (D,the) (N,wine) (FF, .)
(D,the) (N,dove) (V,flies) (P,for) (D,some) (N,flies) (FF, .)
(D,the) (A,last) (N,dove) (V,waits) (A,alone) (FF, .)

a) Draw the graph of the bigram HMM and list all non-zero model parameters that we can obtain via Maximum Likelihood Estimation from the data.
b) Compute the probability of the following sequence according to the previous model:

(D,the) (N,dove) (V,waits) (P,for) (D,some) (A,last) (N,wine) (FF, .)

2. (2 points) Consider Named Entity Recognition and Classification (NERC) for entity types person (PER), location (LOC) and quantities of money (MON).

a) Write generic NERC regular expressions that match named entities occurring in sentences similar to the following ones:

she earns € 1000 a week .
she earns 1000 euros a week .
this house costs one million dollars .
I will spend 25 thousand euros to by a new car .
he is Mr. Smith .
she is Mrs. Smith .
Africa said she will move to London next week .
John London also planned to go to Africa .

in order to get outputs such as the following ones:

she earns [MON € 1000] a week .
she earns [MON 1000 euros] a week .
this house costs [MON one million dollars] .
I will spend [MON 25 thousand euros] to by a new car .
he is [PER Mr. Smith] .
she is [PER Mrs. Smith] .
[PER Africa] said she will move to [LOC London] next week .
[PER John London] also planned to go to [LOC Africa] .

Include the linguistic tags and the resources you require to achieve genericity; we want the patterns to be as more general as possible. Classify your regular expressions as PER, LOC, MON.

b) We are interested in learning a CRF based on BIO model for recognizing and classifying entity types PER, LOC and ORG. Define the template feature involving (t_{i-2}, t_{i-1}, $w_{[1:n]}$, i) for the following description:

the current BIO label is X, the previous one is Y, and one of the two words following the current one is Z

Consider the outputs described above as training examples and write two features derived from that template. How many possible features does that template if the size of the vocabulary is 1000? Justify the answer

3. (3 points) Consider the following PCFG (probabilites for each rule are whown after the rule):

S → NP VP	1.0
NP → DT NBAR	1.0
NBAR → NN	0.7
NBAR → NBAR NBAR	0.3
NP_C → NP C NP	1.0
VP → sleeps	1.0
DT → the	1.0
NN → mechanic	0.1
NN → car	0,2
NN → metal	0.7
C → and	0.6
C → or	0.4

a) Transform this grammar to Chomsky normal form if it is not normalized yet
b) Use the probabilistic CKY algorithm to derive the most likely parse tree for the following sentence.

the metal car mechanic sleeps

Provide the dynamic table of the process. Which is the resulting parse tree and its likelihood?

4. (3 points) A European company wants to benefit from the online comments written by their Spanish and French clients. Some of the comments are opinions about the products of the company. Each opinion refers to a unique product.

The company wants an application able to classify negative opinions with respect to the company's products.

a) What NLP steps and resources are required by the application if the company can provide us with a large corpus and the names of their products?
b) What if the company can only provide us with just a little corpus and the names of their products?
c) Explain the advantages and disadvantages of (a) and (b)