IHLT Exam – 2021

DO NOT USE THIS PAPER TO PROVIDE YOUR ANSWERS
(answers without justification or with a wrong one will be considered wrong answers).

1. (4 points) Given the following subset of rules included in a PCFG

[1] S → NP VP (1.0)
[2] NP → DT NN (0.4)
[3] NP → DT NNS (0.3)
[4] NP → Groucho (0.01)
[5,6] NP → shot (0.03) | elephant (0.04)
[7] NP → NP PP (0.2)
[8] PP → IN NP (1.0)
[9] VP → VP PP (0.5)
[10] VP → VP NP (0.4)
[11] VP → shot (0.03)
[12,13] NN → shot (0.02) | elephant (0.03)
[14] NNS → pajamas (0.02)
[15,16] DT → an (0.2) | his (0.1)
[17] IN → in (0.1)

and the following input sentence:

"Groucho shot an elephant in his pajamas"

a) Apply CKY algorithm and provide the complete, resulting dynamic table. For each component of the table, provide all the information required by the algorithm.
b) Provide the resulting parse tree and its probability. The parse tree must be justified by your answer in (a), if not, it will be considered as wrong answer.
c) Would the CKY result change if one of the following rules was added? Justify your answers briefly.

(1) S → VP (0.3)
(2) NP → NP NP (0.02)

2. (3 points) Suppose you are wanted to build a correct morphological analyzer for English.

a) Answer *Correct/Incorrect* to the following proposals. Justify your answers briefly.

(1) Use Finite State Automata (FSA) combined with a POS tagger.
(2) Use FSA combined with a list of word forms with their corresponding analyses.
(3) Use FSA.
(4) Use a list of word forms with their corresponding analyses.
(5) Finite State Transducers (FST) are more effective than the previous proposals.

b) Given the following forms of verb *take* with their corresponding morphological analyses:

form	analysis
taking	take+VBG
took	take+VBD
takes	take+VBZ

1. Provide the expressions corresponding to the surface level, the intermediate level and the lexical level of a FST for each form. Identify clearly each expression with its respective level.
2. Draw the intermediate FST for those forms.
3. Draw the lexical FST for those forms.
4. How are both FSTs combined to produce the result?

3. (3 points) CRFs can be successfully applied to Named Entity Recognition and Classification (NERC) as well as to Noun-Phrase Recognition (NPR). Suppose we have a vocabulary of 100 words and we want to recognize names of person (PER), organization (ORG), location (LOC) and others (OTH) using BIO notation. (Justify your answers briefly)

a) Which of the following feature templates are incorrect for learning a CRF model for NERC? Which for NPR? Which for both?

$f_{1,a}$: 1 if current word is *a* ; 0 otherwise

$f_{2,a,b}$: 1 if current state is *a* and previous state is *b* ; 0 otherwise

$f_{3,a,b}$: 1 if current state is *a* and 4th previous state is *b* ; 0 otherwise

$f_{4,a,b}$: 1 if 2nd previous word is *a* and previous state is *b* ; 0 otherwise

$f_{5,a}$: 1 if next word is inside a noun-phrase and current state is *a* ; 0 otherwise

b) How many feature functions result from each template in section (a) for NERC task?

c) Suppose the following feature template for learning the NERC model. How do you compute function *semantically_similar(w_1,w_2)* to be productive?

$f_{6,a}$: 1 if *semantically_similar(*current word*,"place")* and current state is a ; 0 otherwise