Cairo University
Faculty of Computers and Information

Bar Code

إسم الطالب

رقم الطالب

Final Exam

Department: Computer Sciences

Course Title: Natural Language Processing
Course Code: CS462
Semester: Spring 2016/2017
Instructors: Dr. Hanaa Bayomi
Date: 4 Jun. 2017
Exam Duration: 2 Hours

Question	Mark	Signature
One		
Two		
Three		
Four		
Five		
Six		
Seven		
Eight		
Nine		
Total Marks		

60

Total Marks in Writing: _____

Attempt ALL Questions:

Q1) you are given the following partial linguistic resources: (10 Marks)

S -> NP VP	(0.9)
NP -> Det N	(0.2)
NP -> N	(0.3)
NP -> NP PP	(0.4)
X -> V NP	(1.0)
VP -> X PP	(0.1)
VP -> V NP	(0.5)
PP -> P NP	(0.8)

And

The	Det	0.1
cat	N	0.2
ran	V	0.7
home	N	0.3
from	P	0.6
garden	N	0.4

a) **What do these resources represent? How are they called?**

b) **Parse the sentence *"the cat ran home from the garden"* using the CYK algorithm.**

c) Draw the parse tree(s) for all possible analyses of the sentence *"the cat ran home from the garden"*.

d) **What is the most probable parse for the former sentence? Justify your answer.**

Q2) Write a transducer for the <u>K insertion spelling rule</u> in English with specifying the surface . (6 Marks)

Name	Description of Rule	Example
K insertion	verbs ending with vowel + -c add -k	panic/panicked

Q3) Imagine that you have a collection of 500 documents and humans have assigned the correct shallow parsing labels for every sentence. Let's call this your "annotated corpus" for shallow parsing. Suppose you have developed a machine learning (ML) classifier that performs shallow parsing, and you decide to evaluate its performance using cross-validation on the annotated corpus. (10 Marks)

a) If you use 5 folds (partitions) for the cross-validation, how many documents will your ML classifier are given for training in each of the 5 experiments?

b) If you use 10 folds (partitions) for the cross-validation, how many documents will your ML classifier be given for training in each of the 10 experiments?

c) If you only had 30 documents in your annotated corpus. Would it be better to use a small number of folds (say, 3 folds) or a large number of folds (say, 10 folds)?

d) If you only had 30 documents in your annotated corpus, would it be better to split your corpus into a single training set and a single test set, or to use cross-validation?

e) If your ML classifier is overly aggressive and labels nearly every word as being in a noun phrase, will its recall or precision be higher for noun phrase labels?

Q4) (16 Marks)

a) "A was salmon outer the does", "It was a nice sunny afternoon", and "I Thomas at mice not the spoon".

What is wrong with such sentences? NLP techniques of what level might allow the system to select the correct one(s)? What would be the required resources?

b) Write a regular expression that accepts any word matched with the pattern " أنستفعل", "أستفعال "

c) **Suppose that we have REgEx ' [Dd][Ee]?[Aa]?[Rr] ' and input string 'Dear Dr.Ashraf, ' How many strings will be matched.**

d) Specify the type of ambiguity in each of the following sentence

Sentence	Type of Ambiguity
1. ***John saw the man on the mountain with a telescope.***	
2. قال علي انه نجح	
3. مدير البنك الجديد	
4. I walk beside the bank	

e) Draw a finite state transducer from bitstrings to bitstrings, which doubles all 1's that are followed by a 0. This means that it should translate 1<u>1</u>00<u>1</u>0011 to 1<u>11</u>00<u>11</u>0011, and 1<u>1</u>001<u>1</u>00 to 1<u>11</u>001<u>11</u>00.

f) For each of the following word determine the number of morphemes and the type of each one [(bounded / free) (derivational / inflectional) (root/stem)]

1) Writers

2) Nationalizations

Q5) Consider the grammar below: **(6 Marks)**

$$
\begin{aligned}
\mathrm{NP} &\rightarrow \mathrm{art\ NP1} \\
\mathrm{NP} &\rightarrow \mathrm{ppro\ NP1} \\
\mathrm{NP1} &\rightarrow \mathrm{num\ NP1} \\
\mathrm{NP1} &\rightarrow \mathrm{NP2} \\
\mathrm{NP2} &\rightarrow \mathrm{adj\ NP2} \\
\mathrm{NP2} &\rightarrow \mathrm{adj\ NP3} \\
\mathrm{NP3} &\rightarrow \mathrm{noun\ NP3} \\
\mathrm{NP3} &\rightarrow \mathrm{noun}
\end{aligned}
$$

State whether the grammar *CAN* or *CAN'T* produce each part-of-speech tag sequence below.

(a) art num noun

(b) art num num noun

(c) art adj noun

(d) art noun

(e) art num adj noun

(f) art num num adj adj noun noun

(g) art adj num noun

(h) art adj adj noun noun

(i) art ppro num noun

(j) ppro num num adj noun noun

(k) ppro noun noun noun

(l) ppro adj adj adj

Q6) (7 Marks)

a) **Briefly describe the specific objectives of the morphological module in the general perspective of automated Natural Language Processing.**

b) **What is the problem addressed by a Part-of-Speech (PoS) tagger? Why isn't it trivial?**

c) **What is the difference between open and closed word class with examples?**

d) **What is the difference between Homonym and Polysemy ?**

Q7) Compare between bottom up and top down parse? (5 Marks)

With My Best Wishes