NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2016/ 2017

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1
Exam(s)	Computer Science – Data Analytics
Module Code(s)	CT5101
Module(s)	Natural Language Processing
Paper No.	1
Repeat Paper	No
External Examiner(s)	Professor Liam Maguire
Internal Examiner(s)	Dr. Michael Schukat *Dr. Paul Buitelaar Dr. John McCrae Dr. Ian Wood Dr. Mihael Arcan

Instructions: Answer all questions. There are 4 sections; each section is worth 25 marks (100 marks total). **Use a separate answer book for each section answered**.

Duration	2 hours
No. of Pages	5
Discipline(s)	Engineering and Information Technology
Course Co-ordinator(s)	Dr. Conor Hayes

Requirements:

Release in Exam Venue	Yes	X	No	
MCQ	Yes		No	X
Handout	None			
Statistical/ Log Tables	None			
Cambridge Tables	None			
Graph Paper	None			
Log Graph Paper	None			
Other Materials	None			
Graphic material in colour	Yes		No	X

CT5101 Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section I: Tagging and Parsing

Instructions: Provide answers for questions 1A, 1B, 1C and 1D

Question 1A 10 Marks

Recall that the Hidden Markov Model is given by the following formula:

$$P(x_1, \dots, x_n, y_1, \dots, y_n) = \prod_{i=1}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

And consider we have the following probability tables:

$P(y\|y')$	$y = N$	$y = V$	$y = A$
$y' = Start$	0.7	0.2	0.1
$y' = N$	0.4	0.5	0.1
$y' = V$	0.5	0.3	0.2
$y' = A$	0.8	0.1	0.1

$P(x\|y)$	$y = N$	$y = V$	$y = A$
$x = I$	0.3	0.1	0.1
$x = like$	0.1	0.7	0.2
$x = Roman$	0.2	0.1	0.6
$x = Numerals$	0.4	0.1	0.1

Given that x_1, x_2, x_3, x_4 = ”I”, “like”, “Roman”, “Numerals”, which part-of-speech tagging is more likely:

- y_0, y_1, y_2, y_3, y_4 = Start, N, V, A, N
- y_0, y_1, y_2, y_3, y_4 = Start, N, A, A, N

Question 1B 5 Marks

Consider the probabilistic context-free grammar as follows

$P(S \rightarrow NP\ VP) = 1.0$
$P(NP \rightarrow N) = 0.7$
$P(NP \rightarrow A\ N) = 0.3$
$P(VP \rightarrow V\ NP) = 0.8$
$P(VP \rightarrow V) = 0.2$

Which of the two part-of-speech taggings (ignoring the Start tag) given in Question 1A has a non-zero probability under this grammar? Show the parse tree that produces this probability.

Question 1C 10 Marks

Briefly describe two disadvantages of probabilistic context-free grammars.

Section 2: Machine Learning

Instructions: Provide answers for questions 2A, 2B, 2C

Question 2A 5 Marks

Briefly explain why it is important to have separate data for training and testing supervised classification models.

Question 2B 10 Marks

Construct a bag of words vector for each of the following 4 sentences. You should use normalization on lower/upper case.

He caught the bus on the way home.
He liked walking home, but buses are usually faster.
The bus was very full and an accident made the bus very slow.
Walking would have been faster.

Question 2C 10 Marks

Below is the table of results from a sentiment analysis classifier applied to a collection of labelled test sentences. Calculate the **precision**, **recall** and **F1** scores of this classifier for the labels +1 and -1, as indicated by these results.

Sentence	**Predicted Label**	**True Label**
I absolutely love it!	+1	+1
Love is overrated!	+1	-1
This is a book about love.	+1	0
Oh! And the lens is double coated! Why would you bother?	0	-1
It's just sad)-:	-1	-1
Mmmmm!! Another one please! (-:	+1	+1
The boy stood on the deck.	0	0
Over and over again! Why me!?	0	-1
Wow! The fool was terrible and we won!	-1	+1
He was so happy that I failed the exam!!	+1	-1

Section 3: Machine Translation

Instructions: Provide answers for questions 3A, 3B, and 3C

Question 3A 10 Marks

Given the translation table below, compute the translation probabilities for the following 4 translations of the German sentence *das Haus ist klein:*

translation1: *the house is small*
translation2: *the house is little*
translation3: *small house the is*
translation4: *the*

f = "das"		***f = "Haus"***		***f = "ist"***		***f = "klein"***	
e	**P(e|f)**	**e**	**P(e|f)**	**e**	**P(e|f)**	**e**	**P(e|f)**
the	0.7	house	0.8	is	0.8	small	0.4
that	0.15	building	0.16	's	0.16	little	0.4
which	0.075	home	0.02	exists	0.02	short	0.1
who	0.05	household	0.015	has	0.015	minor	0.06
this	0.025	shell	0.005	are	0.005	petty	0.04

Question 3B 10 Marks

Consider the following system output and reference translation:

Reference: *The large dog chased the man across the street.*
System: *The big dog chases a man across the street.*

Determine the precision for unigrams, bigrams, 3-grams and 4-grams, and compute the BLEU score.

Question 3C 5 Marks

What change would you expect in the BLEU score, if information about synonyms (e.g. *big* and *large*) is taken into account?

Section 4: Information Extraction

Instructions: Provide answers for questions 4A and 4B

Consider the following 6 sentences about company acquisitions:

LinkedIn announced its acquisition of Lynda.com.
Facebook is close to the acquisition of Pebbles Interfaces.
Facebook is buying a search engine called TheFind.
Apple revealed its acquisition of LearnSprout.
Apple's acquisition of Emotient gives it access to emotion recognition technology.
Siemens is assessing a possible acquisition of wind turbine maker Gamesa.

Question 4A 15 Marks

Describe briefly the 5 basic text processing steps of an information extraction system and give one example of the output for each of these steps when applied to one of the sentences provided above.

Question 4B 10 Marks

What is the Precision, Recall and F-score of an information extraction system that has learnt the "*X * acquisition of * Y*" pattern, where X and Y are company names, applied to the 6 sentences above. Explain how you have derived your answer.

END