NUI Galway
OÉ Gaillimh

Semester 2 Examinations 2018/ 2019

Course Instance Code(s)	1CSD1, 1CSD2, 1SPE1
Exam(s)	MSc in Computer Science (Data Analytics)
Module Code(s)	CT5120
Module(s)	Introduction to Natural Language Processing
Paper No.	1
Repeat Paper	No
External Examiner(s)	Professor Pier Luca Lanzi
Internal Examiner(s)	Dr. Michael Madden *Dr. Paul Buitelaar Dr. John McCrae

Instructions: Answer all parts of 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. Enda Howley

Requirements:

Release in Exam Venue	Yes	☒	No	□
MCQ	Yes	□	No	☒
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

CT5120 Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section 1: Linguistic Foundations

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

Question 1A **10** Marks

Define a constituency (phrase) grammar and lexicon that analyses the following sentence by using the non-terminal symbols ‘S, NP, VP, PP’ and the pre-terminal symbols ‘Det, Noun, Verb, Prep’.

The Taoiseach provided a long answer to questions by TDs.

Question 1B **10** Marks

Draw a constituency (phrase) structure tree and a dependency tree by using the relations ‘nsubj, pobj, amod, det, prep’ for the sentence given in question 1A.

Question 1C **5** Marks

How many types and tokens are there in the sentence given in question 1A?

PTO

Section 2: Language Modelling

Instructions: Provide answers for question 2A, 2B, 2C, 2D and 2E.

Consider the following corpus:

flies fly behind flies then more flies try to fly further behind

Question 2A **5** Marks

State the formula for a bigram language model.

Question 2B **5** Marks

Using a bigram language model without smoothing, calculate the probability of the sentence “flies fly further”. You should use the corpus above to estimate probabilities.

~~**Question 2C** **5** Marks~~

~~Using a bigram language model *with add-one smoothing*, calculate the probability of the sentence “then flies fly further”~~

~~**Question 2D** **5** Marks~~

~~Recall the formula for bigram interpolation~~

~~$p^*(w_n|w_{n-1}) \approx \lambda p(w_n|w_{n-1}) + (1-\lambda)p(w_n)$~~

~~Using a bigram language model *with interpolation* ($\lambda = 0.5$), calculate the probability of the sentence “then flies fly”~~

Question 2E **5** Marks

Why may a language model be used in a machine translation system?

PTO

Section 3: Parsing

Instructions: Provide answers for question 3A, 3B, 3C and 3D

Consider the following probabilistic grammar

N → natural	0.6	NP → A NP	0.1
N → language	0.2	NP → NP NP	0.3
N → processing	0.1	NP → N	0.6
N → works	0.1	VP → V	0.4
A → natural	1.0	VP → V NP	0.4
V → processing	0.1	VP →V NP NP	0.2
V → works	0.9	S → NP VP	0.8
		S → NP	0.2

Question 3A **5** Marks

Describe one ambiguity when applying the above grammar to the sentence “natural language processing works”.

Question 3B **5** Marks

What changes would be necessary to convert the above grammar into Chomsky normal form?

Question 3C **10** Marks

Why should a grammar be in Chomsky normal form when applying the CYK algorithm?

Question 3D **5** Marks

What is a cross-bracketing error and why may it not be important in the example of Q3A?

PTO

Section 4: Distributional Semantics

Instructions: Provide answers for questions 4A and 4B

Consider the following corpus:

A black cat chased the white cat.
The black dog chased the white dog.
A white dog chased the white cat.
A white dog chased the black dog.
The white cat chased a black cat.
The white cat chased a white dog.

Question 4A **15** Marks

Construct a co-occurrence matrix for all types in the corpus, using a context window of two words.

Question 4B **10** Marks

Using Cosine Similarity, compute the distance between:

- *black, white*
- *cat, dog*

END