NUI Galway
OÉ Gaillimh

Semester 1 Examinations 2021/2022

Course Code(s) Instance	1CSD1, 1CSD2, 1SPE1, 1MAO2, 1MAI1
Exam(s)	MSc in Computer Science (Data Analytics), MSc in Computer Science (Artificial Intelligence), MSc in Computer Science (Artificial Intelligence) - Online
Module Code(s)	CT5120, CT5146
Module(s)	Introduction to Natural Language Processing, Introduction to Natural Language Processing - Online
Paper No.	1
Repeat Paper	Yes
External Examiner(s)	Dr John Woodward
Internal Examiner(s)	Dr. Michael Madden *Dr. John McCrae Dr Bharathi Raja Dr Omnia Zayed

Instructions: Answer 4 sections out of 5; each section is worth 25 marks (100 marks total). **Use a separate answer book for each section answered**.

Duration	2 hours
No. of Pages	6
Discipline(s)	Computer Science
Course Co-ordinator(s)	Dr. Frank Glavin Dr. Matthias Nickles Dr. James McDermott

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	[ ]		[X]

Graphic material in colour Yes No

Introduction to Natural Language Processing

Exam Duration: 2 Hours

You must answer 4 of the following sections

Section 1: Text Classification

Question 1A **5** Marks

As we don't know the actual outcome hence using probabilities to model NL able to predict the most likely answer.

Explain in your own words why we use probabilities to model natural language.

Question 1B **10** Marks

Consider the following sentences with sentiment labels.

- This hotel room was great [POS]
- The food was not as great as expected [NEG]
- The pool was great for the kids [POS]
- The kids loved the playground [POS]

$p(pos \mid great) = \frac{C(pos \cap great)}{C(great)} = \frac{2}{3}$.

$p(neg \mid great) = \frac{1}{3}$.

Using Bayes' Law, calculate the probability of the labels POS and NEG given a single feature that considers whether the word 'great' occurs in the text.

Question 1C **10** Marks

What evaluation metrics would you use for a classification problem such as in 1B? Give the formulae for these metrics and explain any advantages or limitations of these metrics.

precision, recall, confusion matrix, F-measures.

PTO

$\frac{TP}{TP+FP}$ $\frac{TP}{TP+FN}$ $\frac{2TP}{TP+TP+FP+FN}$

→ could always achieve 100% by always predicting true/false

F-measure → good for imbalance measure as combined both precision and recall.

Section 2: Linguistic Concept and Parsing

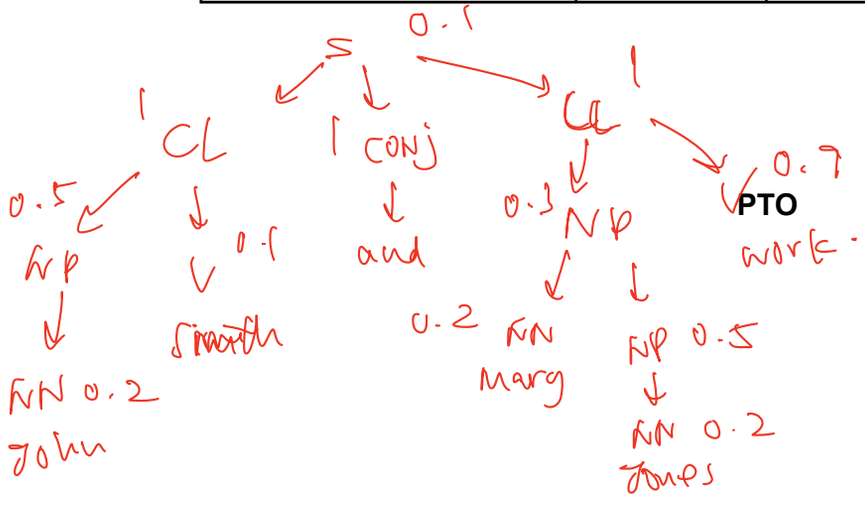
Question 2A **10** Marks

Define constituency (or phrase) grammar and dependency grammars. You should give an example of each. Give **two** reasons why one may be chosen over the other for a particular task.

Question 2B **15** Marks

Consider the probabilistic context-free grammar below. Draw a parse tree and calculate the probability of that parse for the following sentence: "John Smith and Mary Jones work"

Rule	Probability	Rule	Probability
S → NP V	0.9	NN → john	0.2
S → CL CONJ CL	0.1	NN → smith	0.2
CL → NP V	1.0	NN → mary	0.2
NP → NP CONJ NP	0.2	NN → jones	0.2
NP → NN NP	0.3	NN → work	0.2
NP → NN	0.5	V → smith	0.1
CONJ → and	1.0	V → work	0.9

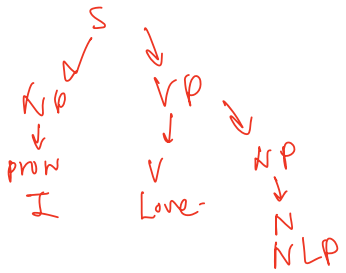


$0.1 \times 1 \times 1 \times 0.5 \times 0.1 \times 0.3 \times 0.9 \times 0.2 \times 0.2 \times 0.5 \times 0.2$

$= 0.0000054$

(2A) - Constituency — to construct a parsing [structure analysis] free for phrase with context-free grammar.

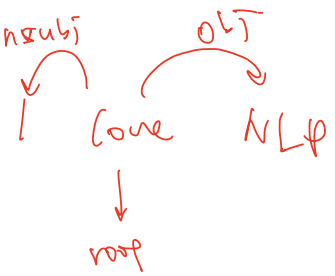
— break sentences to sub-phrases.



e.g. features for classification / assigning semantics roles / mention detection in coreference resolution.

Dependency — to identify [relationship] between words.

— to explore dependencies btwn words in a sentence



e.g. information extraction / Question answering

Section 3: Semantic Analysis

Question 3A **15** Marks

Consider the following sentence:

Priya completed the table with her own ideas

For this sentence, give an example of semantic analysis by means of word sense disambiguation, semantic role labelling and coreference resolution.

Question 3B **5** Marks

Explain how one of the three tasks mentioned above can be solved by a **text classification** approach.

Question 3C **5** Marks

Explain how one of the three tasks mentioned above can be solved by a **tagging** approach.

PTO

(3A) . WSD – Completed + finish + accomplish
– own – a person referring to him/her self.

SRL – Priya (agent), the task (result),
her own ideas (Instrument).

coreference – Priya + her + own.

(3B) · SRL can be completed with text classification:
by identifying the token / phrase is an agent or not.
Similarly for other roles.

(3C) · CR – can be solved by tagging. as tagger is used
to detect the pronouns in sentences.

Section 4: Social Media Analysis

Question 4A **10** Marks

Describe in your own words the main phases of doing social media analysis. Explain the sub-tasks under each phase and highlight the challenges/limitations of each phase.

Question 4B **10** Marks

List at least **five** challenges that could be faced during sentiment analysis on social media data e.g. tweets

Question 4C **5** Marks

List and discuss the main design considerations when building a sentiment analyzer.

PTO

(4A) 1- Data collection — limited data / API / Data privacy regulations.
↳ collect data from social media.

2- Data processing — noise → text normalisation
↓
spelling / abbreviations.
— morphological → lemmetization.

3. Analysis — to decide which model to use.
— performance of models.

(4B) — emoji in tweets
— punctuations
— abbreviations
— spelling.
— short tweets, ambiguity in context

— API limits retrieval.
— not able to access via APIs
— data limited
— implicit sentiment
— indirect sentiment
— irony & sarcasm.
— negation
— informal language

(q) . Considerations ★
- the input (& prediction (sentence / doc)
- the possible output (class / scale / aspect)
- the approach (rule based / supervised / DL)
- the evaluation metrics (ground-truth / prec / recall / acc -)

Section 5: Information Extraction and Vector Space Models

Question 5A **10** Marks

Consider the following text:

Mark Zuckerberg is the founder and CEO of Facebook. He attended Harvard University where he explored different fields including psychology and computer science. With his wife Priscilla Chan, Mark established the Chan Zuckerberg Initiative (CZI) in 2015.

Annotate the sentences above for the named entity types 'person' (PER), 'organization' (ORG) and 'location' (LOC) by the use of the IOB tagging scheme. Explain the reasoning behind your annotations.

Question 5B **5** Marks

Give an example of a hyponym from the text above.

Question 5C **10** Marks

Consider the following text:
The cat lies on the mat.
The dog lies on the floor.
The cat sits near the door.
The dog lies near the door.

with targets 'cat' and 'dog'. Create a vocabulary of context words and a co-occurrence matrix with context N=1.

END

	The	lies	sits
cat	2	1	1
dog	2	2	0