NUI Galway
OÉ Gaillimh

Semester 1 Examinations 2020/2021

Course Instance Code(s)	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	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)	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			
Graphic material in colour	Yes	[ ]	No	[X]

Introduction to Natural Language Processing

Exam Duration: 2 Hours

You must complete Sections 1 to 4

Section 1: Linguistics; Vector Space Model; Semantics

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

Question 1A **5** Marks

Name the morphological process that relates each of the following words pairs:

talk, talking both verbs, inflection
talk, talkative noun/verb and adj, derivation
talk, sweet-talk both nouns, compound

Question 1B **10** Marks

Consider the following grammar and lexicon G. Give all sentences S that can be generated/analysed by G?

G=(N,Σ,P,S)
N: Det, Noun, Verb, NP, VP, S
Σ: *the, men, meet*
P:
S → NP VP
NP → Det Noun
NP → Noun
VP → Verb NP
Det → *the*
Noun → *men*
Verb → *meet*
Start symbol S

Question 1C **5** Marks

Explain what we mean by ‘explicitly defined word senses’.

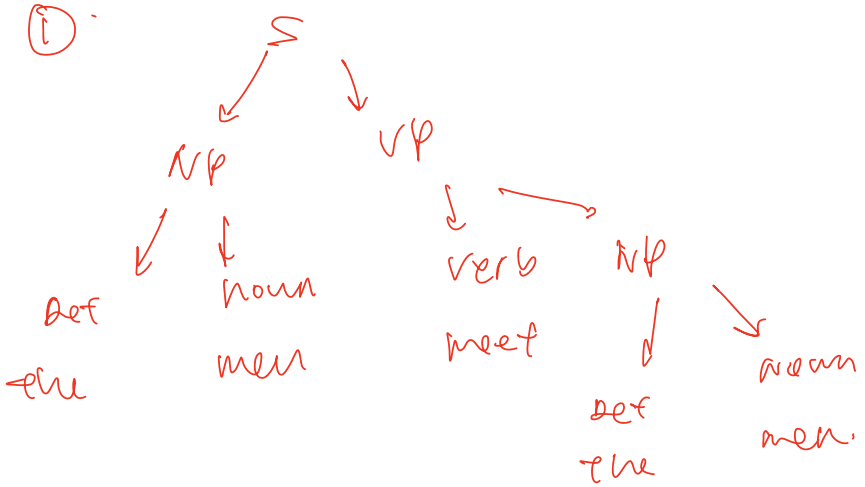
Question 1D **5** Marks

Describe the use of a ‘signature’ in the Lesk algorithm for word sense disambiguation.

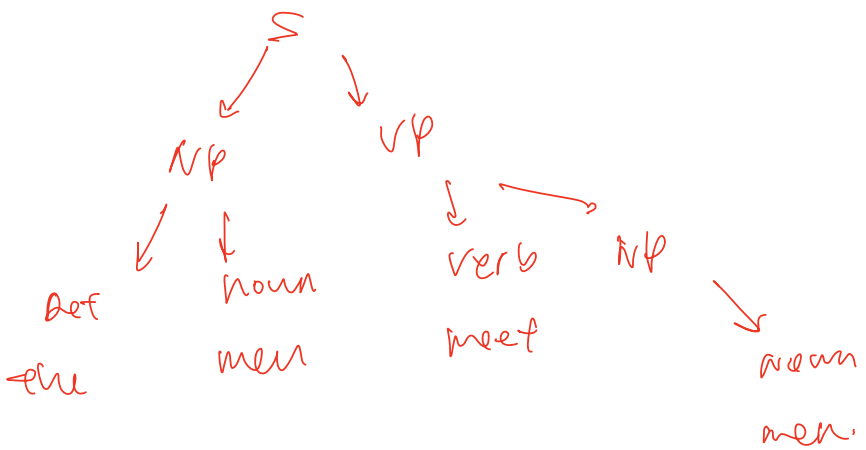
PTO

(18)

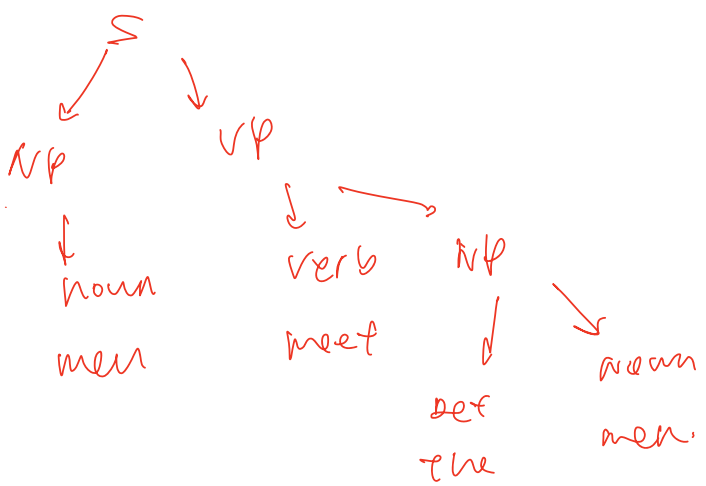
(1)


S
NP
VP
Det
the
noun
men
verb
meet
NP
Det
the
noun
men.

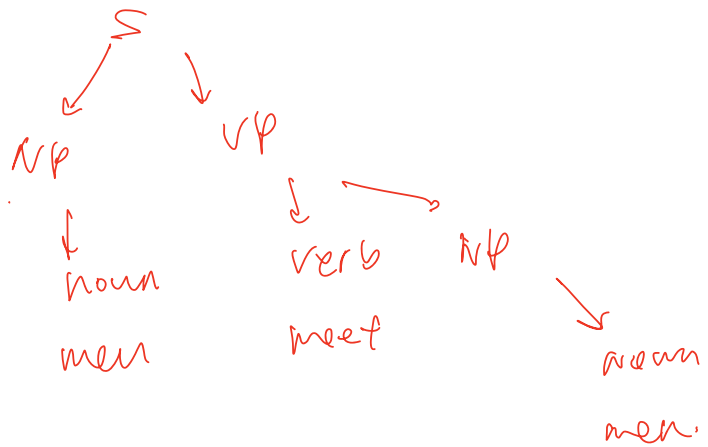
(2)


S
NP
VP
Det
the
noun
men
verb
meet
NP
noun
men.

(3)


S
NP
VP
noun
men
verb
meet
NP
Det
the
noun
men.

(4)



(1C) explicitly defined word senses.

↳ The meaning of the word is explained and associated with definitions.

For example. WordNet is used to disambiguate words by comparing the context of the target & the synsets in WordNet.

(1D) · Lesk- signature is the key words in the dictionary definition in WordNet, which is then used to calculate the overlap between target context and ~~her~~ highest overlap · the true meaning of target word.

WordNet → senses → synsets → dict defi

↓

signatures

Section 2: Language Modeling; Tagging & HMMs; Probabilistic Parsing

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

Question 2A **15** Marks

Consider the following corpus (33 words):

A sailor went to sea sea sea
To see what he could see see see
But all that he could see see see
Was the bottom of the deep blue sea sea sea

Using a bigram language model, calculate the probability tables for all words. State the formula for a bigram language model and use it to calculate the probability of the sentence “he could see what”

Question 2B **5** Marks

Using add-one smoothing and the corpus from 2A, calculate the probability of the sentence “he could see the sea”

Question 2C **5** Marks

Explain how a language model may be used to resolve ambiguities in machine translation

PTO

(2A)

A sailor | Sailor went | went to | to sea | sea sea | sea sea sea | To see | see what | what he | he could | could see | see see | Sure all | all that | that he | was the | the bottom | bottom of | of the | the deep | deep blue | blue sea

$$p(w_1\ w_2\ w_3\ w_4) = p(w_4|w_3) \times p(w_3|w_2) \times p(w_2|w_1) \times p(w_1)$$

$$p(\text{he could see what}) = \frac{1}{5} \times 1 \times 1 \times \frac{2}{33}$$

$$= \frac{1}{5} \times \frac{2}{33} = \frac{2}{165}$$

$$= 0.0121$$

answer = 0 is also correct,
as P(he|Start) = 0

2B $p(\text{the could see the sea}) = \frac{2+1}{33+8} \times \frac{2+1}{2+21} \times \frac{2+1}{2+21} \times \frac{1+1}{5+21} \times \frac{0+1}{2+21}$

$p(w) = \frac{c(w)+1}{N+V}$

$p(w_2|w_1) = \frac{c(w_1 w_2)+1}{c(w_1)+V}$

assume V is unique bigram.

=

bigram add-one = c(bigram) + 1 / (c(prior) + number of unique words (same as unigram)

2C Probabilistic context free grammar may be used to resolve ambiguities as able to output higher probabilities for grammatically correct. parse tree. then machine could translate based on the highest possible parses.

for example : I can fish

highest possibility =

N V N

hence translate to another language with similar parse. instead of NNV.

Section 3: Information Extraction; Knowledge Graphs & Chatbots

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

Question 3A **10** Marks

Identify Hearst patterns in the following Wikipedia text on Ireland and explain for each how you would apply them to extract hyponym relations:

"Irish culture has had a significant influence on other cultures [...]. Alongside mainstream Western culture, a strong indigenous culture exists, as expressed through Gaelic games, Irish music and the Irish language. The island's culture shares many features with that of Great Britain, including the English language, and sports such as association football, rugby, horse racing, and golf."

Question 3B **10** Marks

Using the IOB format, annotate the following sentence from the Wikipedia text on Ireland with entities of type LOCATION. Explain your annotation.

"Geopolitically, Ireland is divided between the Republic of Ireland (officially named Ireland), which covers five-sixths of the island, and Northern Ireland, which is part of the United Kingdom."

Question 3C **5** Marks

Entity linking may use a knowledge graph as context for entity disambiguation, in addition to textual context information. Suggest one method that utilises a knowledge graph for entity linking..

PTO

Section 4: Opinion Mining, Ethics & Data Privacy

Instructions: Provide answers for questions 4A, 4B and 4C

Question 4A **10** Marks

Consider the following sentiment lexicon:

avoid	0.00 POS, 1.00 NEG
awful	0.00 POS, 1.00 NEG
nice	1.00 POS, 0.00 NEG
lovely	1.00 POS, 0.00 NEG
recommend	1.00 POS, 0.00 NEG

What are the Positive (POS) and Negative (NEG) sentiment scores for the following review, using the sentiment lexicon as defined above:

"This is a very nice hotel situated in the center of GALWAY. Lovely bar and restaurant. Nice food, lovely staff. I would highly recommend this hotel."

Question 4B **5** Marks

For the same review as in Question 4A, identify the sentiment aspects.

Question 4C **10** Marks

Describe in your own words NLP aspects of Data Privacy.

END

(4A) $\frac{+}{\text{nice}}$ $\frac{-}{-}$

lovely

nice

lovely

recommend

5 pos $\Rightarrow \frac{5}{26}$ #pos score

0 neg. $\Rightarrow \frac{0}{26}$ Neg score

$f(\frac{5}{26}, \frac{0}{26}) = +ve$

(4B) · Location, bar & restaurant, food, ~~hotel~~ hotel.

(4C) · Data privacy. → involves personal data.
Should be anonymized.
irreversasley.

→ Consent should be taken when data used in ML projects involves personal data.

→ should obey regularizations such as GDPR.