NUI Galway
OÉ Gaillimh

Autumn 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	Yes
External Examiner(s)	Professor Pier Luca Lanzi
Internal Examiner(s)	Dr. Michael Madden *Dr. Paul Buitelaar, Dr. John McCrae

<u>Instructions</u>: 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

<u>Requirements</u>:

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	☒

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 and 1C

Question 1A — **5** Marks

Explain the difference between stemming and lemmatization. Give an example of each.

Question 1B — **10** Marks

Consider the following sentences.

The man met the woman.
The woman met the man.
The man met the woman and the man.
The woman and the man met the woman.

S → NP VP
NP → Det N conj Det N
VP → V NP
NP → Det N

Define the grammar and lexicon G that can be used to generate these sentences by giving N, Sigma and completing the rules P as follows:

G=(N,Σ,P,S)
N: def, n, V, NP, VP, S, conj
Σ: the, man, met, woman, and
P: S → NP VP

Lexical Rules
Det → the
N → man | woman
V → met
conj → and.

Start symbol S

Question 1C — **10** Marks

Consider the following frequency vectors. Using cosine similarity, compute the distributional semantic distance between ‘dog’ and ‘cat’.

cat	*0*	*5*	*5*	*2*	*0*

dog	*0*	*3*	*4*	*2*	*6*

sim (dog, cat)

$$= \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\ \sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$= \frac{0+15+20+4+0}{\sqrt{25+25+4}\ \sqrt{\cancel{0}+16+4+36}}$$

$$= \frac{39}{\sqrt{54}\ \sqrt{65}} = 0.6583.$$ #

PTO

1A Stemming – obtain stem of word

– remove the different ending and keep the shared part.

e.g. learns, learning, learnt

→ learn

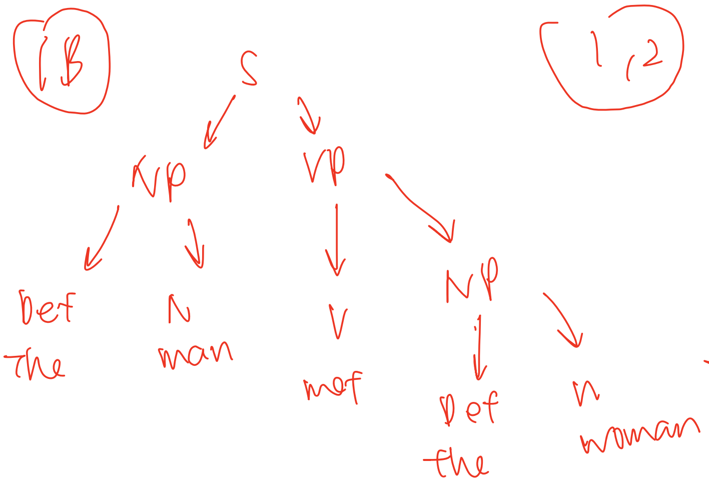
Lemmatization – obtain the root word.

– reduce to original form

e.g. studies, studying, studied

→ study.

1B



1,2

S → NP VP

NP → Det N

VP → V NP

NP → DEt N

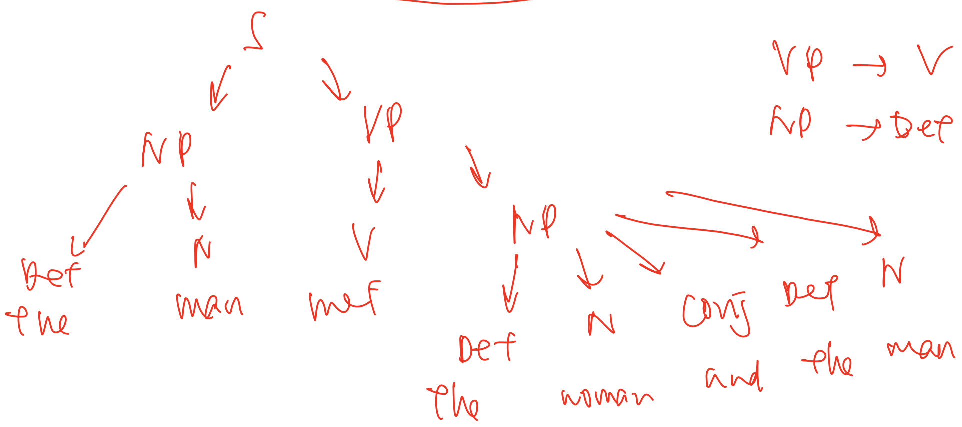
3

S → NP VP

NP → Det N

VP → V NP

NP → Det N conj DetN



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

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

Question 2A **10** Marks

Consider a Hidden Markov Model with the following probabilities (Start designates the start state):

$p(w_i \mid t_i)$	w_i = this	w_i = question	w_i = is	w_i = easy
t_i = N	0.4	0.4	0.1	0.1
t_i = V	0.1	0.2	0.5	0.2
t_i = O	0.4	0.1	0.1	0.4

$p(t_i \mid t_{i-1})$	t_{i-1} = N	t_{i-1} = V	t_{i-1} = O	t_{i-1} = Start
t_i = N	0.2	0.2	0.7	0.1
t_i = V	0.7	0.3	0.1	0.2
t_i = O	0.1	0.5	0.2	0.7

By using the Viterbi algorithm or otherwise, what is the most likely sequence of tags for the text "this question is easy"?

O N V V 0.00165

Question 2B **10** Marks

Given a corpus with a part-of-speech tag values for each word, how would you learn the probabilities for a table such as in question 2A?

Question 2C **5** Marks

How would you modify the Viterbi algorithm in order to produce the probability of the text over all possible combinations of part-of-speech tags?

PTO

2A

~~p(this question is easy, O N V V)~~ = $\frac{0.7 \times 0.7 \times 0.7 \times 0.3 \times}{0.4 \times 0.4 \times 0.5 \times 0.2}$ = 0.00105

~~p(this question is easy, O N V N)~~ = $\frac{0.7 \times 0.7 \times 0.7 \times 0.2 \times}{0.4 \times 0.4 \times 0.5 \times 0.1}$ = 0.00055

~~p(this question is easy, O V V V)~~ = $\frac{0.7 \times 0.1 \times 0.3 \times 0.3 \times}{0.4 \times 0.5 \times 0.2 \times 0.5}$ = 0.000126

~~p(this question is easy, O V V N)~~ = $\frac{0.7 \times 0.1 \times 0.3 \times 0.2}{0.4 \times 0.5 \times 0.2 \times 0.1}$ = 0.000168

2B → Transmission = $p(t_i \mid t_{i-1})$

$= \frac{C(MD, VB)}{C(MD)}$

modal verbs

verb base

↳ compute max likelihood estimate of this transition probability by counting, out # times we see the first tag in a labeled corpus, how often the first tag is followed by the second.

→ Emission = $p(w_i \mid t_i) = \frac{C(\overset{\text{tag}}{MD}, \overset{\text{e.g.}}{will})}{C(MD)}$

↳ probability of a given tag associated with the word.

Section 3: Information Extraction; Knowledge Graphs & Chatbots

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

Question 3A **10** Marks

Explain what Hearst patterns are, how they are used in information extraction and give an example of three different Hearst patterns with a corresponding example sentence for each.

Question 3B **5** Marks

Consider the following positive instance (Source Wikipedia) for the relation Play_For between entities of type FOOTBALLER and TEAM:

[FOOTBALLER Kevin De Bruyne] is a Belgian professional footballer who plays as a midfielder for [TEAM Premier League club Manchester City].

Give a negative instance for this relation and these entity types, not involving negation.

Question 3C **10** Marks

Give all taxonomy elements, terms and term pairs, that can be identified in the following text (Source Wikipedia):

A bank is a financial institution that accepts deposits from the public and creates a demand deposit while simultaneously making loans. A bank borrows money by accepting funds deposited on current accounts, by accepting term deposits, and by issuing debt securities such as banknotes and bonds.

PTO

Section 4: Opinion Mining, Ethics & Data Privacy

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

Question 4A **10** Marks

What is SentiWordNet and how can it be used in sentiment analysis?

Question 4B **5** Marks

For the following review, identify the sentiment aspects:

Exceptional customer service from all staff we couldn't have asked for a warmer welcome. Food is amazing too and such a great location for travelling around Cork.

Question 4C **10** Marks

Describe in your own words NLP aspects of data privacy.

END

4A SentiWord Net → derived from WordNet, adding sentiment polarity on (words in) WordNet synset s

→ able to provide an unsupervised lexicon based approach for sentiment analysis.

4C It is challenging to collect data for NLP tasks as API / existing datasets are often limited. Some of the researchers might come across crawling on webpages but it could violate the data privacy as the source is not authorized hence violating the GDPR regulations.