Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2020-2021
M.Tech (Data Science and Engineering)
End-Semester Test (EC-3 Makup)

Course No. : DSECLZG525
Course Title : Natural Language Processing
Nature of Exam : Open Book
Weightage : 50%

No. of Pages = 4
No. of Questions = 5

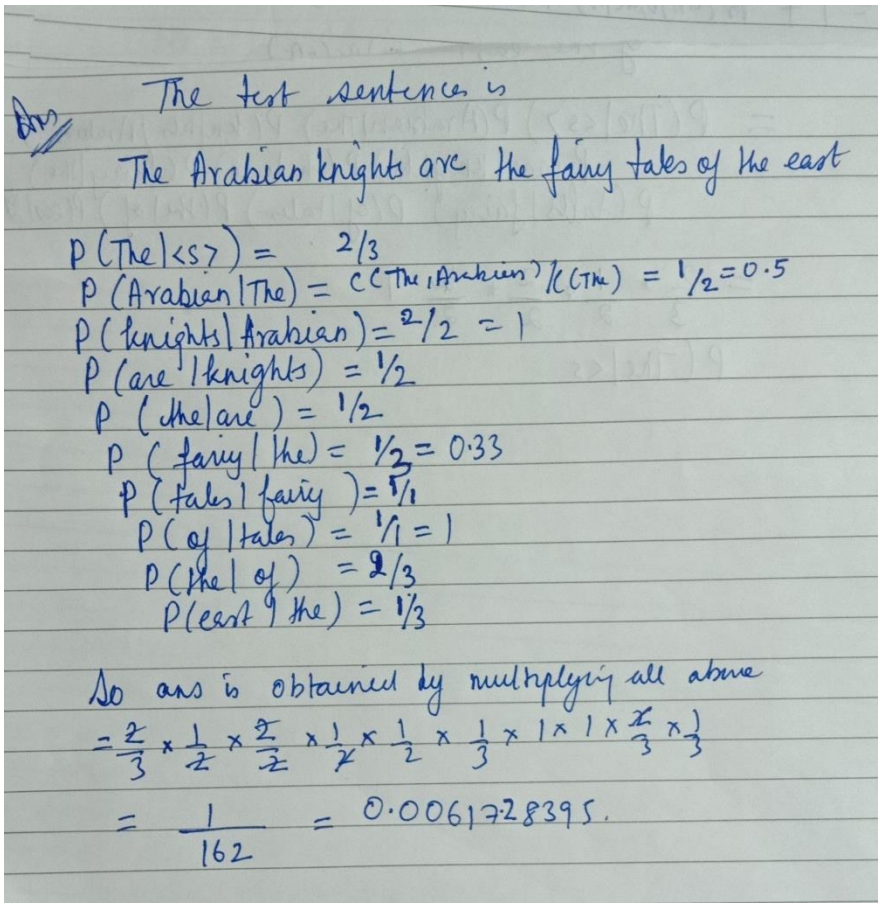
Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1.

a) Consider the training set: (4 marks)

The Arabian knights
These are the fairy tales of the east
The stories of the Arabian knights are translated in many languages

Compute using the bigram model the probability of the sentence. Include start and end symbol in your calculations.
The Arabian knights are the fairy tales of the east



b) Using Penn Tree bank, find the POS tag sequence for the following sentences: [6 Marks]

1. The actor was happy he got a part in a movie even though the part was small. [2 marks]
2. I am full of ambition and hope and charm of life. But I can renounce everything at the time of need [3 marks]
3. When the going gets tough, the tough get going. [1 mark]

Solution

The/DT actor/NN was/VB happy/JJ he/PRP got/VB a/DT part/NN in/IN a/DT movie/NN "even though"/CC the/DT part/NN was/VB small/ADV. [2 marks]

I//PRP am/VB full/JJ of/IN ambition/NN and/CC hope/NN and/CC charm/JJ of/IN life/NN. But/CC I/PRP can/VB renounce/VB everything/JJ at/IN the/DT time/NN of/IN need/NN [3 marks]

When/WDT the/DT going/NN gets/VB tough/RB, the/DT tough/NN get/VB going/RB.[1 mark]

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	*and, but, or*	SYM	symbol	*+,%, &*
CD	cardinal number	*one, two, three*	TO	"to"	*to*
DT	determiner	*a, the*	UH	interjection	*ah, oops*
EX	existential 'there'	*there*	VB	verb, base form	*eat*
FW	foreign word	*mea culpa*	VBD	verb, past tense	*ate*
IN	preposition/sub-conj	*of, in, by*	VBG	verb, gerund	*eating*
JJ	adjective	*yellow*	VBN	verb, past participle	*eaten*
JJR	adj., comparative	*bigger*	VBP	verb, non-3sg pres	*eat*
JJS	adj., superlative	*wildest*	VBZ	verb, 3sg pres	*eats*
LS	list item marker	*1, 2, One*	WDT	wh-determiner	*which, that*
MD	modal	*can, should*	WP	wh-pronoun	*what, who*
NN	noun, sing. or mass	*llama*	WP$	possessive wh-	*whose*
NNS	noun, plural	*llamas*	WRB	wh-adverb	*how, where*
NNP	proper noun, singular	*IBM*	$	dollar sign	*$*
NNPS	proper noun, plural	*Carolinas*	#	pound sign	#
PDT	predeterminer	*all, both*	"	left quote	' or "
POS	possessive ending	*'s*	"	right quote	' or "
PRP	personal pronoun	*I, you, he*	(	left parenthesis	[, (, {, <
PRP$	possessive pronoun	*your, one's*	)	right parenthesis	],), }, >
RB	adverb	*quickly, never*	,	comma	,
RBR	adverb, comparative	*faster*	.	sentence-final punc	. ! ?
RBS	adverb, superlative	*fastest*	:	mid-sentence punc	: ; ... – -
RP	particle	*up, off*			

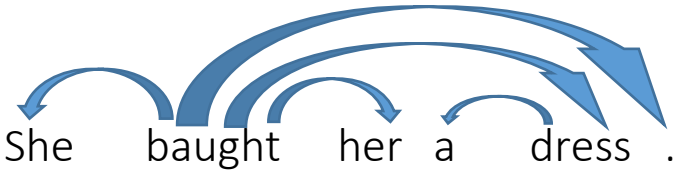
Question 2.

a) Build a parse tree for the sentence "She loves to visit Goa" using Probabilistic Parsing [5marks]

S → NP VP 1.0
VP → V PP 0.4
VP → V NP 0.6
PP → P NP 1.0
NP -> V NP 0.1
NP → NP PP 0.3
NP → N 0.3
N → visit 0.3
V → visit 0.6

N → Goa 0.3
N → She 0.5
V → loves 1
P → to 1
DT→ a 1

a) State the correct sequence of actions that generates the following parse tree of the sentence “She baught her a dress" using Arc-Eager Parsing [5marks]



Solution:
Transitions: SH-LA-SH-RA-SH-LA-RE-RA-RE-RA
Arcs:
She <- baught
baught _> her
a <- dress
baught -> dress
baught -> .

Question 3. Word sense disambiguation and ontology-

b) What are lexical sample task and all word task in word sense disambiguation? How can sources like Wikipedia be used for word sense disambiguation [2 marks]

Solution

What are lexical sample task and all word task in word sense disambiguation?

Lexical sample task and all word task are 2 variants of word sense disambiguation

- Lexical sample task -Small pre-selected set of target words
- All-words task - System is given an all-words entire texts and lexicon with an inventory of senses for each entry. We have to disambiguate every word in the text (or sometimes just every content word).

How can sources like Wikipedia be used for word sense disambiguation

Wikipedia can be used as training data for word sense disambiguation using supervised learning techniques

- Concept is mentioned in a Wikipedia: article text may contain an explicit link to the concept’s Wikipedia page, which is named by a unique identifier (can be used as a sense annotation)
- These sentences can then be added to the training data for a supervised system.

How can WordNet relations be used for word sense disambiguation in following sentences: [3 marks]

1. A bat is not a bird, but a mammal.
2. Jaguar reveals its quickest car ever
3. Raghuram Rajan was the 23rd Governor of the Reserve Bank of India

Solution

Nouns and verbs can be extracted from the sentences. The senses in wordnet can be extracted for these words and senses with close relations can be extacted as correct sense.

1. Bat can be sports bat or mammal. But looking at nouns bat, bird and mammal, correct sense of bat as MAMMAL can be found using WordNet relations.
2. Jaguar can be a car or animal. Looking at nouns Jaguar, correct sense of Jaguar as CAR can be found using WordNet relations.
3. Bank can be river bank or financial bank.: Search senses of nouns Bank,"Raghuram Rajan", Governer. The correct sense of BANK as FINANCIAL sense can be found using WordNet relations.

c) How is Syntactic web different from the Semantic web? What is URI in semantic web ontology? [2 marks]

Syntactic web consist of huge data on net connected by hyperlinks which is rendered by machines but machines cannot process it due to inability to understand the meaning of the content.

The semantic Web identifies a set of technologies, tools, and standards which form the basic building blocks of an infrastructure to support the vision of the Web associated with meaning.

A Universal Resource Identifier (URI) is a formatted string that serves as a means of identifying abstract or physical resource. A URI can be further classified as a locator, a name, or both. Every resource is identified with unique URI in ontology.

Develop an OWL ontology using the following for animal kingdom for classes like carnivorous, herbivorous and omnivorous. Use following Property characteristics, restrictions and Class expressions [3 marks]

- inverseOf
- domain
- range
- Cardinality
- disjointWith
- subClassOf

```
<rdfs:Class rdf:ID="Carnivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Herbivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Omnivorous">
  <rdfs:subClassOf rdf:resource="#Animal"/>
</rdfs:Class>

<rdfs:Class rdf:ID="Carnivorous">
  <owl:disjointWith rdf:resource="#Herbivorous"/>
</rdfs:Class>
<owl:Restriction>
   <owl:onProperty rdf:resource="#hasLegs" />
   <owl:cardinality
 rdf:datatype="&xsd;nonNegativeInteger">4</owl:cardinality>
 </owl:Restriction>
 <owl:ObjectProperty rdf:ID="Eats">
   <rdfs:domain rdf:resource="#Carnivorous"/>
```

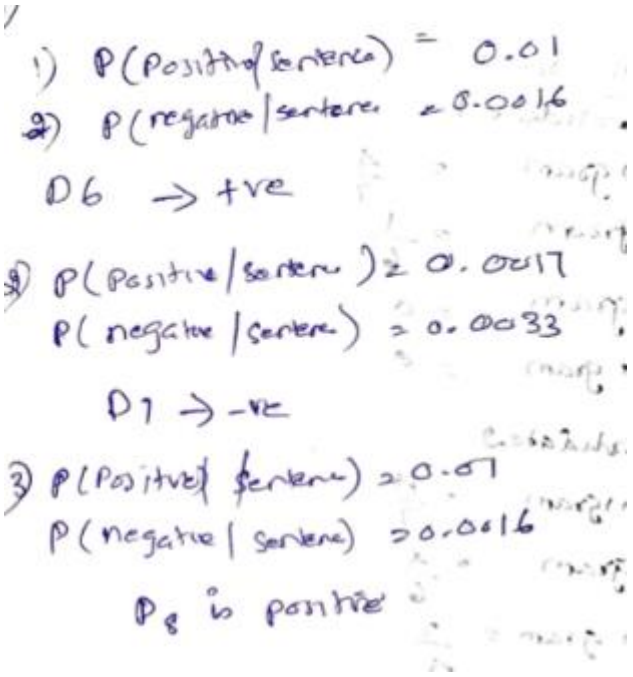
```
  <rdfs:range  rdf:resource="#Animal"/>
</owl:ObjectProperty>
```

Question 4.

a) In this modern age where the internet is growing rapidly, the existence of the internet can make it easier for tourist to find information about hotels. Tourists usually tell the experience during the hotel by writing reviews on the internet. Hence many hotel's reviews are found on the internet. With the availability of reviews on the internet with large numbers, tourists can't understand all the reviews they read whether they contain positive or negative opinions. It takes a sentiment analysis to quickly detect if the reviews is a positive or negative reviews. Using the Multinomial Naïve Bayes Classifier method find out that the given hotel reviews are positive or negative.

D1	The hotel is clean and great	Positive
D2	The hotel owner is very helpful	Positive
D3	Overall Aston Hotel's experience was great	Positive
D4	The condition of the hotel was very bad	Negative
D5	A HORRIBLE EXPERIENCE FOR ONE WEEK	Negative
D6	The hotel view was great	?
D7	My holiday experience stay in usa so horrible	?
D8	Overall the hotel in aston very clean and great	?

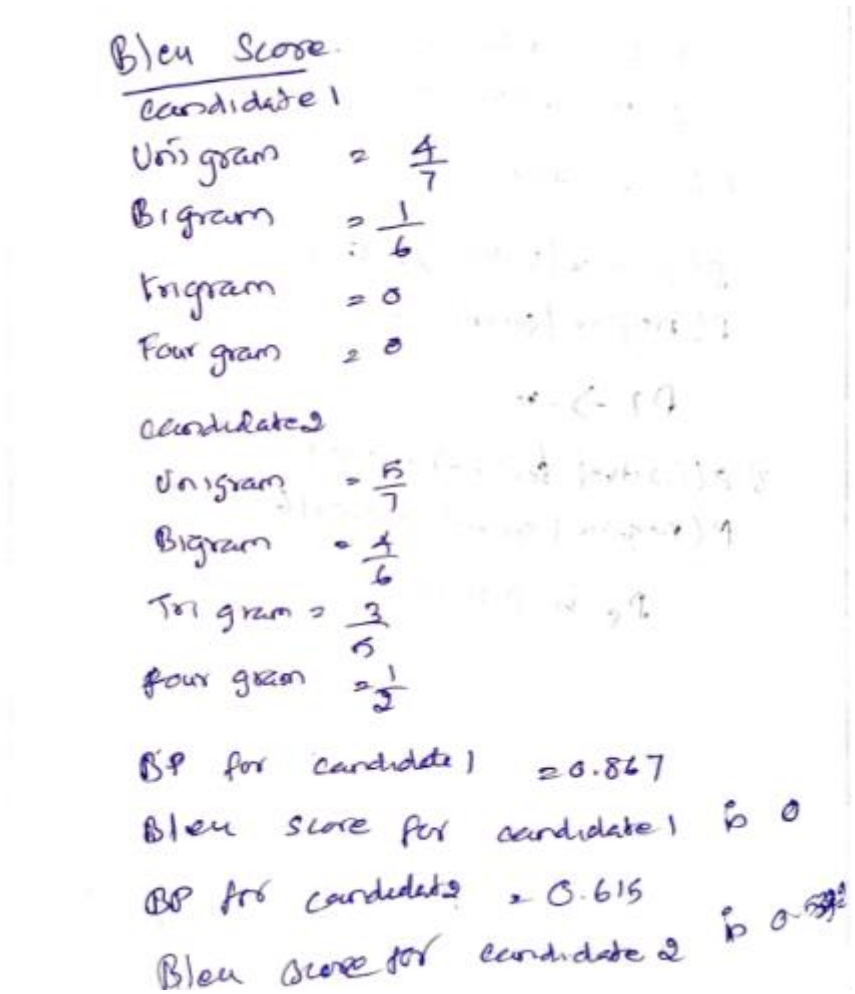
Soln :



b. Compute the BLEU score for the below translations (candidate1, candidate2).Consider 1gram,2 gram,3 gram,4 gram and Brevity-Penalty for calculating BLUE score .

Reference: The teacher arrived late because of the traffic

Candidate 1: The teacher was late due to the traffic
Candidate 2: A teacher arrived late because of transportation

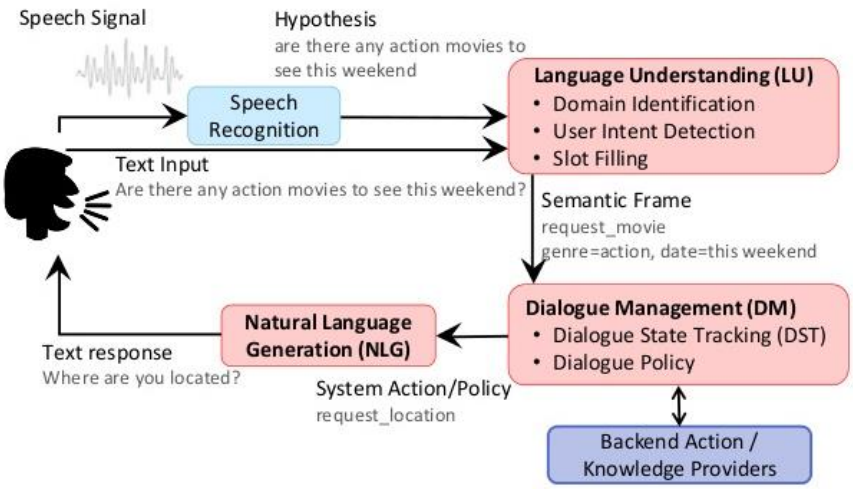


1. Identify the type of the dialogue agent from the following conversation. Also explain the working of that type of dialogue agent. (5 marks)

   System: How may I help you?
   User: I want to book dinner at a restaurant that serves good Chinese food
   System: Berkos is a restaurant that serves good Chinese food.
   User: ok
   System: Should I book a dinner?
   User: yes, please
   System: ok.

   Soln: Above is the task oriented dialogue system

The diagram shows the working of the task oriented dialogue system. The student should explain briefly each step as mentioned in the slides



2. Find the intent, domain and slots for the following: (2 marks)
   Book an appointment on 12th Feb 2021 at 10 am for a ECG Test.
   SOLN:
   DOMAIN: Medical
   INTENT: Book an Appointment

   Slots

   - Services: ECG TEST
   - Date: 12th Feb 2021
   - Time: 10 AM

3. In a collection of 10000 document, the following words occur in the following number of documents: (3 marks)
   Oasis occurs in 400 documents, Place occurs in 3500 documents, Desert occurs in 800 documents, Water occurs in 800 documents, Comes occur in 800 documents
   Beneath occurs in 200 documents, Ground occurs in 900 documents
   Calculate TF-IDF term vector for the following document:
   Oasis Place Desert Water Comes Beneath Ground Place

Term	(TF) Term freq.	IDF	TF*IDF
Oasis	1/8	log(10000/400)	0.1747
Place	2/8	log(10000/3500)	0.11398
Desert	1/8	log(10000/800)	0.137114
Water	1/8	log(10000/800)	0.137114
Cenes	1/8	log(10000/800)	0.137114
Beneath	1/8	log(10000/200)	0.212371
Ground	1/8	log(10000/900)	0.13072

TF-IDF vector (0.1747, 0.11398, 0.137114, 0.137114, 0.137114, 0.212371, 0.13072).