Det ->the

Verb->Book

Prep->on

Noun->flight|airasia

Solution:

S->VP

VP->Verb NP

VP->V NP

NP->NP PP

NP->Det Noun

PP->Prep Noun

Verb	Det	Noun	preposition	Noun

0 Book 1 the 2 flight 3 on 4 airasia 5

VP->V.NP

NP-> Det.Noun NP->DetNoun. PP->Prep Noun.

NP->NP. PP

3. Given the grammar and lexicon below find the probabilty of the best parse tree using PCFG for the below sentence [5 marks]

S: I book the dinner flight

Grammar		Lexicon
$S \rightarrow NP\ VP$	[.80]	$Det \rightarrow that$ [.10] | a [.30] | the [.60]
$S \rightarrow Aux\ NP\ VP$	[.15]	$Noun \rightarrow book$ [.10] | $flight$ [.30]
$S \rightarrow VP$	[.05]	| $meal$ [.15] | $money$ [.05]
$NP \rightarrow Pronoun$	[.35]	| $flights$ [.40] | $dinner$ [.10]
$NP \rightarrow Proper\text{-}Noun$	[.30]	$Verb \rightarrow book$ [.30] | $include$ [.30]
$NP \rightarrow Det\ Nominal$	[.20]	| $prefer;$ [.40]
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I$ [.40] | she [.05]
$Nominal \rightarrow Noun$	[.75]	| me [.15] | you [.40]
$Nominal \rightarrow Nominal\ Noun$	[.20]	$Proper\text{-}Noun \rightarrow Houston$ [.60]
$Nominal \rightarrow Nominal\ PP$	[.05]	| NWA [.40]
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does$ [.60] | can [40]
$VP \rightarrow Verb\ NP$	[.20]	$Preposition \rightarrow from$ [.30] | to [.30]
$VP \rightarrow Verb\ NP\ PP$	[.10]	| on [.20] | $near$ [.15]
$VP \rightarrow Verb\ PP$	[.15]	| $through$ [.05]
$VP \rightarrow Verb\ NP\ NP$	[.05]	
$VP \rightarrow VP\ PP$	[.15]	
$PP \rightarrow Preposition\ NP$	[1.0]	

Solution:

Attached the screenshot

Language model

1. Consider the training set:

   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
   Soln:

Ans// The test sentence is

The Arabian knights are the fairy tales of the east

$P(\text{The} \mid \text{<s>}) = 2/3$

$P(\text{Arabian} \mid \text{The}) = C(\text{The, Arabian}) / C(\text{The}) = 1/2 = 0.5$

$P(\text{knights} \mid \text{Arabian}) = 2/2 = 1$

$P(\text{are} \mid \text{knights}) = 1/2$

$P(\text{the} \mid \text{are}) = 1/2$

$P(\text{fairy} \mid \text{the}) = 1/3 = 0.33$

$P(\text{tales} \mid \text{fairy}) = 1/1$

$P(\text{of} \mid \text{tales}) = 1/1 = 1$

$P(\text{the} \mid \text{of}) = 2/3$

$P(\text{east} \mid \text{the}) = 1/3$

So ans is obtained by multiplying all above

$$= \frac{2}{3} \times \frac{1}{\cancel{2}} \times \frac{\cancel{2}}{\cancel{2}} \times \frac{1}{\cancel{2}} \times \frac{1}{2} \times \frac{1}{3} \times 1 \times 1 \times \frac{\cancel{2}}{3} \times \frac{1}{3}$$

$$= \frac{1}{162} = 0.0061728395.$$

---

2. You are an English class teacher and to make your course interesting you hold a write like Shakespeare competition where you ask each student to write a play in the style of Shakespeare. You try to score the plays automatically by using a trigram model where the probability distribution for the trigrams is

calculated using all of Shakespeare's plays as the corpus. While scoring the student plays however, you find that the data was not enough and most sentences in the student's plays have a score of 0. What options do you have to come out of your predicament if you still want to score the plays automatically?

Ans:

The options are the same as when we have inadequate language models. We back-off to simpler models e.g. bigram and unigram models and use smoothing to compute probabilities where the corpus does not have instances of the relevant n-gram. Another option is to use a weighted linear combination of multiple n-gram models for different values of n (e.g. n=1, 2, 3).

Part of speech tagging and HMM

1. 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]

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*			

Soln:

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]

2.

Part of speech tagging using Hidden Markov Model (HMM) – M2

The following three sentences (S1, S2 and S3) and their corresponding tag sequences (T1, T2 and T3) are given as training data for implementing HMM. **Answer questions A-D.**

S1: John cut the paper .	S2: Mary asked for a hair cut .	S3: Sharon asked for a pay cut .
T1: N V D N STOP	T2: N V I P N N STOP	T3: N V I P N N STOP

What will be size of Emission and Tag translation matrices? **Note: Include the start and the stop symbols and assume we are working with a bigram model.**
What will be the Emission Probability P(asked | V) ?
What will be the Tag Translation Probability P(I|V) ?
If a brute force approach is employed to find the tags for the test sentence "**I had a deep cut**.", how many possible tag sequences need to be evaluated?

Parsing

1. 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 → 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

2) Give the correct sequence of arc eager parsing operations for the given sentence [2marks]

Solution:

[Root,]	[The lazy cat slept]	[]
[Root,The]	[lazy cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[Shift]
[Root ,the ,Lazy]	[cat slept]	[LA]
[Root the]	[cat slept]	[LA]
[Root,Cat]	[slept]	[SH]
[Root,]	[slept]	[LA]
[Root,Slept]	[]	[RA]

SH,SH,LA,LA,SH,LA,RA

3. Given the grammar and lexicon below dervie the parse tree using top down parsing method for the sentence [3 marks]

S :The guy ate pizza

S->NP VP

VP->VNP

NP->Det N

N->pizza

N->guy ,Det ->the

V->ate

<u>Soln:</u>

1The 2 guy 3 ate 4 the 5pizza 6

State	Backup State	Action
1.((S) 1)		
2.((NP VP) 1)		
3.(DT N VP) 1)		matches the
4.((N VP) 2)		matches guy
5.((VP)3)		

6.((V NP) 3) matches ate

7.((Det N) 4)
matches the

8.((N))5
matches pizza

9.()

4. Given the grammar and lexicon below show the final chart for the following sentence after applying the bottom up chart parser.[5 marks]

S: Book the flight on airasia

S->VP

VP->V NP

NP->NP PP

NP->Det Noun

PP ->Prep Noun

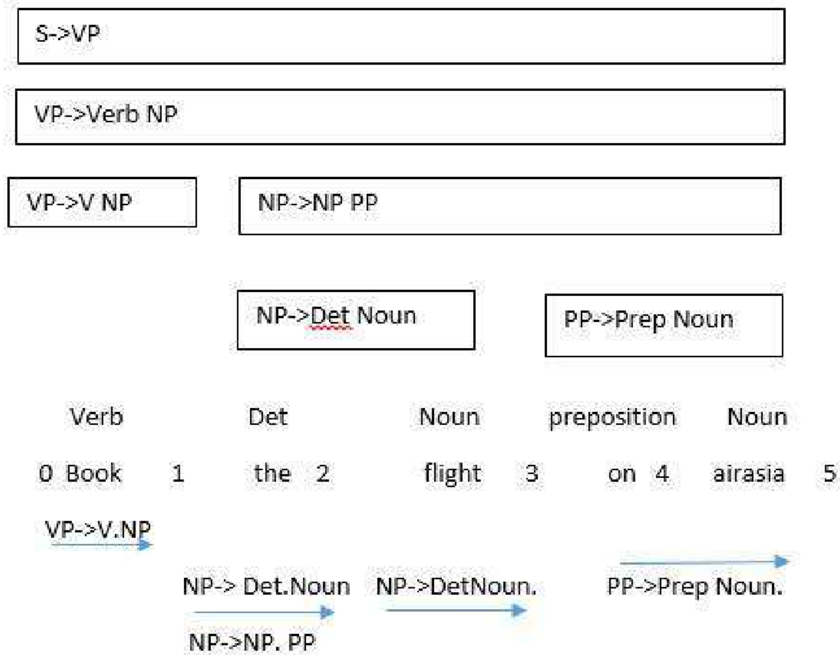
Det ->the

Verb->Book

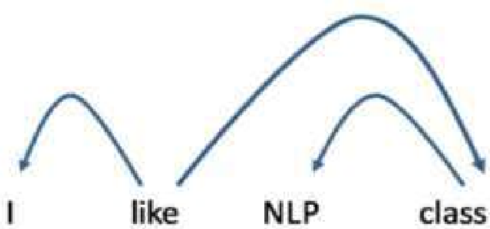
Prep->on

Noun->flight|airasia

Soln:



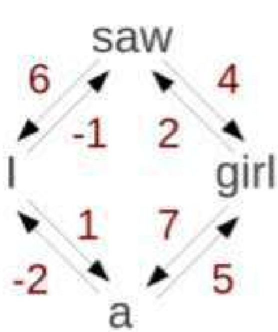
5. Correct sequence of actions that generates the following parse tree of the sentence “I like NLP class" using Arc-Eager Parsing is [5marks]



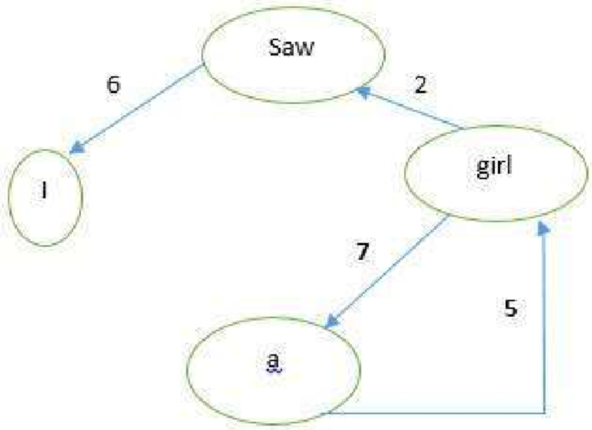
Soln:

SH->LA->RA->SH->LA->RA->RE->RE->RE

6. In the below weighted graph, the edge weights between girl-saw and a-girl in the maximum spanning tree are: [5]



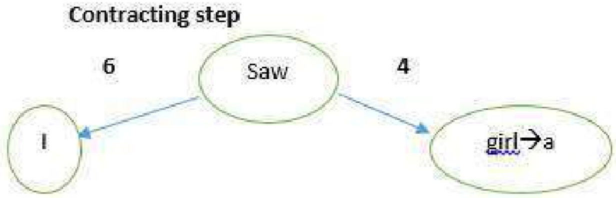
Soln:



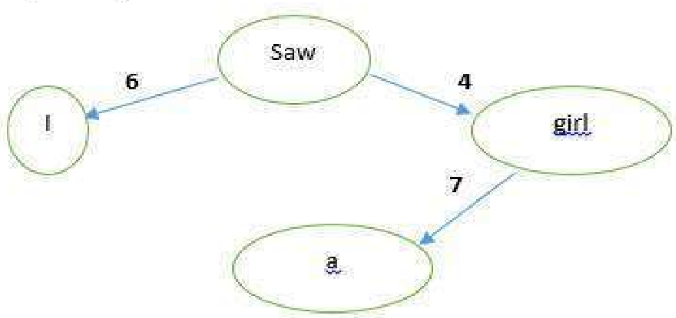
A → girl → forms a cycle

As per Chu Liu Edmond algorithm we can remove the cycle by contracting and expansion

Contracting step



Expanding



Because saw ->girl -> a gives more weightage than a->girl->saw

So the final weight for saw ->girl is 4 and a-> girl is 7

.

7. Given a treebank how would you determine probabilities for the grammar rules given in Question 1 (for use with a basic PCFG parser)?

Let's take the VP rule. There are three VP rules. I would count the total number of VP rules in the

Treebank. Then, for each rule, I would count the number of times that rule occurs and divide by the total number of VP rules. That would yield the probability for each rule. I would follow a similar procedure for each rule where the same non-terminal appeared on the left-hand side.

8.

1. Given the grammar and lexicon below (which is the same as that of question 5), show one possible **top-down derivation** for the sentence:

   Run the Detroit marathon

S → NP VP	Det → the
S → VP	Noun → run — marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	

Answer

S → VP
S → Verb NP
S → Verb Det NP
S → run Det NP
S → run the Proper-Noun Noun
S → run the Detroit Noun
S → run the Detroit marathon

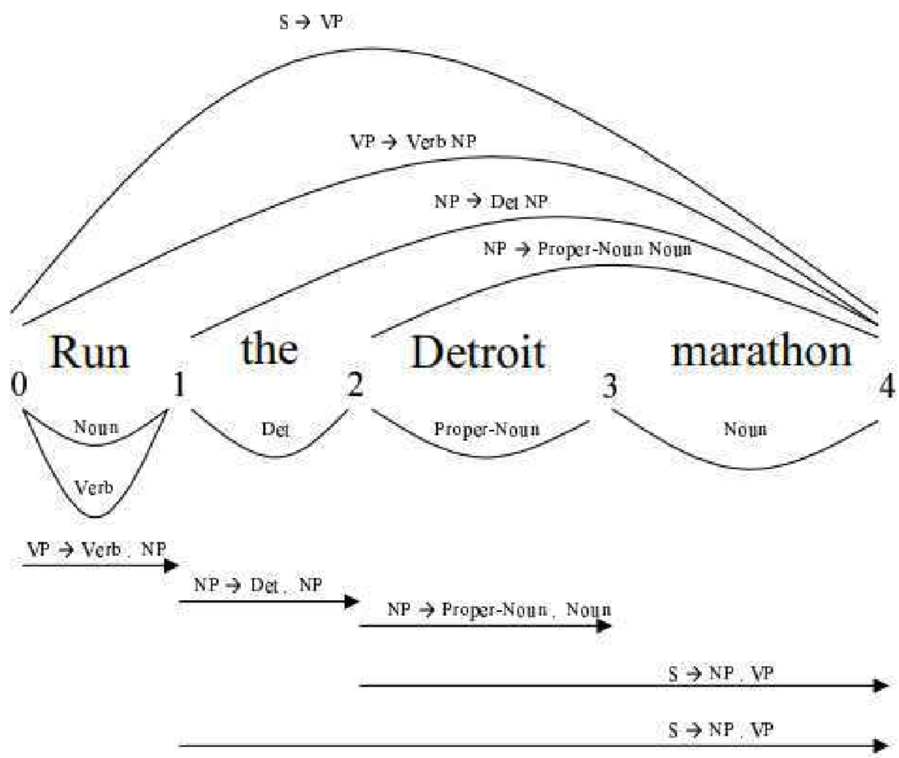
9.

Given the grammar and lexicon below, show the **final chart** for the following sentence after applying the bottom-up chart parser from class:

Run the Detroit marathon

Remember that the final chart contains all edges added during the parsing process. You may use either the notation from class (i.e. nodes/links) or the notation from the book to depict the chart.

S → NP VP	Det → the
S → VP	Noun → run | marathon
NP → Det NP	Verb → run
NP → Proper-Noun Noun	Proper-Noun → Detroit
VP → Verb NP	


S → VP
VP → Verb NP
NP → Det NP
NP → Proper-Noun Noun
Run
the
Detroit
marathon
0
1
2
3
4
Noun
Verb
Det
Proper-Noun
Noun
VP → Verb . NP
NP → Det . NP
NP → Proper-Noun . Noun
S → NP . VP
S → NP . VP

____________________**********________________________

Sentiment analysis

1.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 Positive	?

Soln:

After smoothing

word	P(word|positive)	[illegible]
		2/32
Hotel	4/26	1/32
clean	3/26	1/32
great	3/26	1/28
owner	3/26	2/32
very	2/26	4/32
Helpfull	2/26	1/32
overall	2/26	1/32
Asian	2/26	2/32
experience	2/26	2/32
condition	1/26	2/32
Bad	1/26	2/32
Horrible	1/26	2/32
one	1/26	2/32
week	1/26	3/32

1) P(Positive|sentence) = 0.01

2) P(negative|sentence = 0.0016

D6 → +ve

3) P(Positive|sentence) = 0.0017

P(negative|sentence) = 0.0033

D7 → -ve

3) P(Positive|sentence) = 0.01

P(negative|sentence) = 0.0016

D8 is positive

2. Given the following documents and their sentiment polarities

Document	Sentiment words	Polarity
D1	Good, Enjoy, Good	Positive
D2	Poor, Unpleasant	Negative
D3	Enjoy ,Wonderful	Positive
D4	Good, Lovely	Positive
D5	Good, Poor, Rude	Negative
D6	Good ,Wonderful	?

Determine the sentiment polarity of document D6 using the multinomial naïve Bayes classification (with add1 smoothing) approach. Show your step in detail.

Solution:

P (Positive) =3/5

P (Negative) =2/5

P (Good /Positive) =3+1/7+7=4/14 P (Good/Negative) =1+1/5+7=2/12

P (Wonderful/Positive) =1+1/7+7 =2/14 P (Wonderful/negative) =0+1/5+7=1/12

<u>For the document 6</u>

P (Positive/Good, Wonderful) = 4/14*2/14*3/5

=0.29*0.14*0.6

=0.024

P (Negative/ Good, Wonderful) =2/12*1/12*2/5

=0.16*0.083*0.4

=0.005

Sentiment polarity of document D6 is Positive

3. For sentiment analysis of twitter data, Which classifier did you choose among SVM and Naïve bayes and why? Justify your answer.[3m]

Ans:

Option -1 mark

Justification (2marks) since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

b**)** ***I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought it. She also thought the phone was too expensive, and wanted me to return it to the shop.*** **What are the problems associated with this type of sentiment analysis?[3Marks]**

Ans:

Identification of Implicit aspects.

Multiple sentiments for same opinion phrase for different aspects

Association of corresponding sentiments to aspects in a multi-aspect review

4. How is sentiment calculated or scored? [2M]

5. Compare Rule based approach and machine learning approach in sentiment analysis [5].

6.

Given a list of positive and negative seed sentiment words and Table-1 providing the details of word occurrence and co-occurrences:

positive seeds (Pwords): good, nice, excellent, positive, fortunate, correct, superior

negative seeds (Nwords): bad, nasty, poor, negative, unfortunate, wrong, inferior

Compute the polarity of the phrase "excellent and outstanding".

Word 1	Word 2	Count of word 1	Count of word 2	count of co-occurrences
excellent	outstanding	5578	2749	1384
poor	outstanding	283891	3293296	3347

Table-1

Machine translation

1. 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

Soln:

Bleu Score

Candidate 1

Unigram $= \frac{4}{7}$

Bigram $= \frac{1}{6}$

Trigram $= 0$

Four gram $= 0$

Candidate 2

Unigram $= \frac{5}{7}$

Bigram $= \frac{4}{6}$

Tri gram $= \frac{3}{5}$

four gram $= \frac{1}{2}$

BP for candidate 1 $= 0.867$

Bleu score for candidate 1 is 0

BP for candidate 2 $= 0.615$

Bleu score for candidate 2 is 0.[illegible]

2. 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 NASA Opportunity rover is battling a massive dust storm on Mars.

Candidate 1: The Opportunity rover is combating a big sandstorm on Mars.

Candidate 2: A NASA rover is fighting a massive storm on Mars.

Soln:

Metric	Candidate 1	Candidate 2
precision1 (1gram)	8/11	9/11
precision2 (2gram)	4/10	5/10
precision3 (3gram)	2/9	2/9
precision4 (4gram)	0/8	1/8
Brevity-Penalty	0.83	0.83
BLEU-Score	0.0	0.27

3. Compute the BLEU score for the machine transaltion output .

Assume 1gram ,2 gram and *brevity penalty*

German: Ich bin zur Zeit nicht im Büro

Reference English: I am currently out of the office

MT English: I am currently not in the office

Unigram

$$p_1 = \frac{1+1+1+0+0+1+1}{7} = \frac{5}{7} = 0.71$$

Bigram

English: (I am), (am currently), (currently out), (out, of), (of the), (the office)

MT English: (I am), (am currently), (currently not), (not in), (in the), (the office)

P1=1+1+1+1+1+1/6 =4/6=0.67

brevity penalty =1

*Final BLUE Score =1*0.71*0.67=0.48*

4. Compute the BLEU score for the MT output.

(Assume 1gram,2 gram.3 gram.4 gram and brevity penality)

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

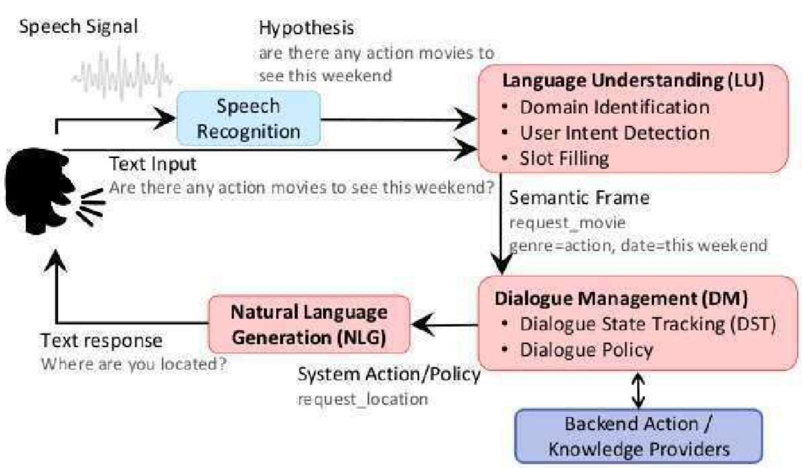
Question and answer system and Chat bot

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

Soln

	(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
Comes	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).

Word sense disambiguation and ontology

1) 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).

2. 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

Ans:

• 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.

3. 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 word net can be extracted for these words and senses with close relations can be extracted 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.

4. Consider below three types of movie reviews and convert it into Bag of words model:

- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good
Review 1	1	1	1	1	1	1	1	0	0	0	0
Review 2	1	1	2	0	0	1	1	0	1	0	0
Review 3	1	1	1	0	0	0	1	0	0	1	1

2.

- *the cat sat*
- *the cat sat in the hat*
- *the cat with the hat*

.

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1