Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

NLP MSE-2

Q1(a) Discuss Lesk algorithm

Ans Lesk algorithm is an overlap approach used for word sense disambiguation.
Steps involved are—

1) Preprocessing—
Tokenize the input text into individual words.

2) Remove stop words & punctuation marks to focus on content words.

2) Retrieve definitions
Retrieve dictionary definitions for the ambiguous word from a lexical resource like wordnet

3) Calculate overlap
- Compare the definitions of the ambiguous word with the definitions of its content words.
- Calculate the overlap between the glosses (definitions) of the ambiguous word & the glosses of its content words.
- The overlap can be calculated using various techniques, such as counting words or using semantic similarity measures

4) Select the best sense
- Choose the sense with the highest overlap as the disambiguated sense for the ambiguous

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

word.

- The sense with the highest overlap indicates the most relevant meaning in given context

Eg:

On Burning coal we get ash

Sense Bag: It contains the words in the definition of a candidate sense of the ambiguous word.

Context Bag - It contains the words in the definition of each sense of each context word.

Ash

Sense 1 - Tree of the olive family.

Sense 2 - Solid residue left when combustible material is thoroughly burnt or oxidized

Sense 3 - To convert into ash.

Coal

Sense 1 - A piece of glowing carbon or burnt wood.

Sense 2 - Charcoal

Sense 3 A black solid combustible substance

In this case Sense 2 of ash will be winner ∵ it has maximum overlapping words.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Q1(b)	Discuss the application of WordNet & BabbleNet Lexical database
	Ans	Application of WordNet
		- Synonym & Antonym Extraction
		- Word sense disambiguation.
		- Text classification & sentiment analysis
		- Query expansion in information Retrieval
		- Ontology Building & knowledge Representation
		Application of BabbleNet
		- Multilingual WSD
		- Cross language Information Retrieval
		- Multilingual knowledge Representation
		- Name Entity Recognition.
		- Semantic Text similarity
		- Cross language Plagiarism Detection.
	Q1(c)	Explain how a supervised learning algorithm can be applied for word sense disambiguation
	Ans	Supervised algorithm assumes that context can supply enough evidence to disambiguate words on its own.
		- It involves training a model using a labeled dataset of word senses. The model is then used to disambiguate the sense of a target word in a new text.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	–	Decision List – A decision list is a set of rules that assign to a target word based on context in which it appears
	–	Neural Network – Neural Network such as feedforward, recurrent RNN & transformer NN are used to model the context-sense relationship.
	–	Support vector Machine – used feature vectors. In the context of WSD, words & their context are represented as features like part of speech, word embedding, linguistic information. The SVM learns to find a hyperplane that best separate the data points representing the different sense.
	–	Naive Bayes – It is probabilistic algorithm that uses Bayes theorem to classify text into predefined categories.
	–	Decision Tree – It is flowchart-like structure in which an internal node represents feature (or attribute) the branch represents a decision rule & each leaf node represents the outcome.

Total Marks of Question no. | Examiner 1

Space for Marks	Question No.	START WRITING HERE

Q2(a) Convert the following CFG grammar into CNF

$S \to A \mid cB \mid c$

$A \to Bc \mid b$

$B \to cAA \mid \varepsilon$

Sol'n CNF

$V \to V_1 V_2$

$V \to T$

Remove Null Production

$S \to A \mid cB \mid c$

$A \to Bc \mid b \mid c$

$B \to cAA$

Remove Unit Production

$S \to Bc \mid b \mid c \mid cB$

$A \to Bc \mid b \mid c$

$B \to cAA$

Let $X \to AA$ $Y \to c$

$S \to BY \mid b \mid c \mid YB$

$A \to BY \mid b \mid c$

$B \to YX$

//

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Q2(b)	What are the challenges of NLP & how to resolve them
	Ans	1) Ambiguity Lexical (word having multiple meaning) syntactic (different ways to parse) semantic (same word interpreted differently) Solution - WSD - Use contextual clues & statistical models to determine the correct sense of a word in a given context. - Deep learning models like transformer (BERT) which takes the entire sentence or paragraph into account can reduce ambiguity. 2) Understand Context & Pragmatics Challenge - understanding context, especially in complex texts or dialogs remains difficult. - Context embedding - Models like BERT, GPT4 generate dynamic word embedding that change depending on the surrounding words. - Reinforcement learning - Systems can be trained using real-world interaction to better interpret pragmatics, intent & tone.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	3)	Handling Long dependencies challenge - Understanding long-range relationships b/w words. Sol^n - Transformers - BERT, GPT.
	4)	Data sparsity challenge - for low resource or specialized domain there must be labeled data for training NLP models Sol^n - Transfer learning, data augmentation
	5)	Domain adaptation Legal, medical applications trained on general dataset is a challenge Sol^n - fine tuning of dataset
	6)	Cultural & linguistic Diversity Sol^n Multilingual model - MBERT Crowdsourcing & low resource model

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	Q2(c)	Discuss various POS Tagging methods
	Ans.	POS Tagging methods
	1)	Rule Based POS Tagging
	2)	Stochastic POS Tagging
	3)	Transformation Based Tagging
		Rule Based POS Tagging
		- They use dictionary or lexicon for getting possible tags for tagging each word.
		- If the word has more than one possible tag then it use handwritten rules to identify the correct tag.
		- The can was rusted
		Here can is either noun or modal verb But since the is Determinant so can shd be noun.
		Stochastic POS Tagging
		- The model that includes frequency or probability is called stochastic.
		- Stochastic taggers generally resolve tagging ambiguity by using a training corpus to compute the probability of a given word having a given tag in a given ~~any~~ context.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
		It uses two n taggers —
	1)	Word frequency approach stochastic tagger disambiguate the words based on the probability that a word occurs with a particular tag.
	2)	Tag sequence probability The tagger calculate the probability of a given sequence of tags occuring
		– Transformation Based tagging A hybrid approach. – Like rule-based taggers – this tagger is based on rules. – Also automatically induced from hand-tagged data.

Total Marks of Question no.		Examiner 1

Space for Marks	Question No.	START WRITING HERE

Q3(a) Using HMM pos tagging tag "The park is a Book"

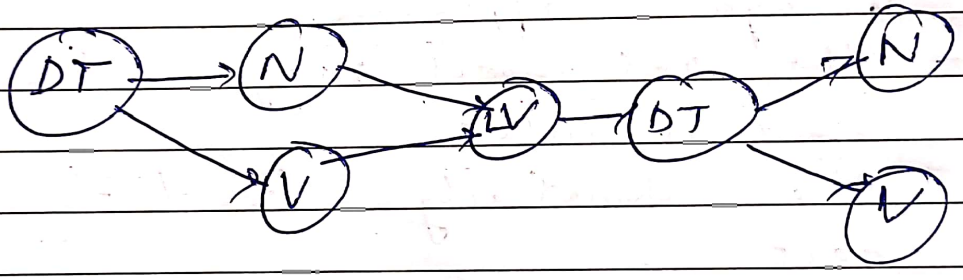
Corpus

<s> Book a car </s>
<s> Park the car </s>
<s> The book is in the car </s>
<s> The car is is a park </s>

Sol'n

Step 1

The Park is a Book



<s> Book a Car </s>
V DT N

<s> park the car </s>
V DT N

<s> The Book is in the car </s>
DT N LV IN DT N

<s> The car is in a park </s>
DT N LV IN DT N

DT - Determinant
N - Noun
LV - Linking verb
IN - Preposition
V - Verb.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

Step2: Emission Probability Matrix

	N	V	LV	IN	DT
The	0	0	0	0	4/6
park	1/6	1/2	0	0	0
is	0	0	2/2	0	0
a	0	0	0	0	2/6
Book	1/6	1/2	0	0	0

Transition Matrix

	N	V	LV	IN	DT	<F>
<S>	0	2/4	0	0	2/4	0
N	0	0	2/6	0	0	4/6
V	0	0	0	0	2/4	0
LV	0	0	0	2/2	0	0
IN	0	0	0	0	2/2	0
DT	6/6	0	0	0	0	0

Step3

(1) The is DT

$$P(The|N) = P(N|<S>) \times P(The|N)$$
$$= \frac{2}{4} \times \frac{4}{6} = \frac{1}{6}$$

P_c $P(The|N) = \frac{1}{6}$

(2) Park is now

$$P(Parks|N) = P(N|DT) \times P(Park|N)$$
$$= \frac{1}{6} \times \frac{6}{6} \times \frac{1}{6} = \frac{1}{36}$$

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE		
		Book as verb	is is LV	a is DT
		$= P(\text{Book}/V) \times P(V/\text{is}) \times P(\text{is}/\text{DT})$	$= \frac{1}{36} \times \frac{2}{2} \times \frac{2}{6}$	$= \frac{1}{108} \times \frac{1}{324} \times \frac{2}{6}$
		$= \frac{1}{2} \times \frac{1}{1000} \times \frac{1}{6}$	$= \frac{1}{108}$	$= \frac{1}{1944}$
		$= \frac{1}{6000}$		
		∴ Book is Noun	is is IV	a is DT
		Book as Noun	Book as verb	∴ Book is Noun.
		$\frac{1}{324} \times \frac{1}{6} \times \frac{6}{6}$	$\frac{1}{324} \times \frac{1}{2} \times \frac{1}{6000}$	

Q3(b) Explain –

(1) Homonymy – Same word diff meaning
Bank – financial bank
Bank – river bank

(2) Homophone – same sound diff meaning
Right ✗ Write

(3) Polysemy – multiple related meaning
By Bank

(4) Synonym – Same meaning different word
Sofa / couch

(5) Antonym – opposite
light / dark.

(6) Hyponym
Car is hyponym of animal vehicle

(7) Hypernym – Vehicle is hypernym of car