

## NLP – Unit 4 (Information Retrieval using NLP) – END-SEM PYQ Answers

May-June 2023

### Q3(a) — Information Retrieval System in NLP

[4 Marks]

- **Information Retrieval (IR):** Process of obtaining relevant information from a large collection (corpus, web, database) in response to a query.
- **IR System Components:**
  - 1. Document Collection — corpus of documents to search
  - 2. Indexer — preprocesses and indexes documents (tokenization, stemming, stopword removal)
  - 3. Query Processor — interprets and reformulates user queries
  - 4. Retrieval Model — ranks documents by relevance (Boolean, VSM, BM25)
  - 5. Results Ranker — presents top-k ranked documents to user
- **Role of NLP in IR:**
  - Query understanding: POS tagging, NER, intent detection
  - Text normalization: stemming, lemmatization, stopword removal
  - Semantic search: word embeddings, query expansion
  - Passage ranking: contextual models like BERT for re-ranking

*Note: Example: Google Search is an IR system. When you type 'Python tutorial', NLP processes the query, retrieves relevant web pages, and ranks them.*

### Q3(b) — Named Entity Recognition (NER) and Evaluation Metrics

[8 Marks]

- **NER Definition:** NER identifies and classifies named entities in text into predefined categories.
- **Common Entity Types:**
  - PERSON — 'Elon Musk', 'Albert Einstein'
  - ORGANIZATION — 'Google', 'SPPU'
  - LOCATION — 'Mumbai', 'India'
  - DATE/TIME — 'May 2023', '9 PM'
  - MONEY — '\$500', '₹10,000'
  - MISC — product names, events, etc.
- **NER System Pipeline:**
  - 1. Tokenization → 2. POS Tagging → 3. Feature Extraction → 4. NER Model → 5. Entity Tags
- **Tagging Schemes:** IOB (Inside-Outside-Beginning): B-PER, I-PER, O

- **Evaluation Metrics:**

Metric	Formula	Meaning
Precision	$TP / (TP + FP)$	Of entities predicted, how many are correct?
Recall	$TP / (TP + FN)$	Of actual entities, how many are found?
F1-Score	$2 \times P \times R / (P + R)$	Harmonic mean of Precision and Recall
Accuracy	$(TP+TN) / \text{Total}$	Overall correctness (less used for NER)

- **Exact Match vs Partial Match:** Strict: entire span must match. Partial: only some tokens match.

*Note: F1-Score is the primary metric for NER evaluation. High precision means fewer false positives. High recall means fewer entities missed.*

### Q3(c) — Cross-Lingual Information Retrieval (CLIR)

[6 Marks]

- **CLIR Definition:** Retrieving documents in a different language than the query language.
- **Example:** User queries in English → retrieve relevant documents in Hindi/French/German.
- **Approaches to CLIR:**
  - 1. Query Translation: Translate query from source to target language, then search
  - 2. Document Translation: Translate all target documents to query language
  - 3. Corpus-Based Methods: Use parallel corpora/bilingual dictionaries for mapping
  - 4. Cross-Lingual Word Embeddings: Map words across languages into shared vector space
- **Challenges:**
  - Ambiguity in machine translation (multiple meanings of a word)
  - OOV (Out-of-Vocabulary) words and proper nouns
  - Grammar differences between languages
  - Resource scarcity for low-resource languages
- **Real-World Example:** A researcher types 'climate change effects' in English → CLIR retrieves French scientific papers on 'effets du changement climatique'.

*Note: CLIR is crucial for multilingual web search, legal research across countries, and cross-national scientific collaboration.*

### Q4(a) — Vector Space Model (VSM) in Information Retrieval

[6 Marks]

- **VSM Concept:** Represents documents and queries as vectors in a multi-dimensional term space.
- **How it Works:**
  - Each unique term = one dimension in the vector space
  - Each document = a vector of term weights (TF-IDF values)

- Query = also represented as a vector
- Similarity = Cosine similarity between document and query vectors

- **Cosine Similarity Formula:**

$$\text{sim}(d, q) = (d \cdot q) / (|d| \times |q|) = \sum(d_i \times q_i) / \sqrt{\sum d_i^2} \times \sqrt{\sum q_i^2} \quad \text{Range: 0 (no similarity) to 1 (identical direction)}$$

- **VSM Example:**

- Doc1: [TF-IDF values for each term in vocabulary]
- Query: [TF-IDF values for query terms, 0 for absent terms]
- Compute cosine similarity of Doc1 with Query → rank documents by score

Strengths	Weaknesses
Simple and efficient	Ignores word order
Handles partial matches	No semantic understanding
Ranking with similarity scores	Sparse vectors for large vocab
Easy to implement	Synonyms treated as different terms

#### Q4(b) — Entity Extraction and Relation Extraction

[8 Marks]

- **Entity Extraction:** Identifying and extracting named entities (persons, organizations, locations, etc.) from text.
- **Example — Entity Extraction:**
  - Text: 'Sundar Pichai is the CEO of Google, headquartered in California.'
  - Extracted: [PERSON: Sundar Pichai], [ORG: Google], [LOC: California]
- **Entity Extraction Methods:**
  - Rule-Based: Regular expressions and handcrafted rules (e.g., capitalized words)
  - ML-Based: CRF, HMM, LSTM with IOB tagging scheme
  - Deep Learning: BERT-based NER models (fine-tuned transformers)
- **Relation Extraction:** Identifying semantic relationships between entities in text.
- **Example — Relation Extraction:**
  - Text: 'Marie Curie was born in Poland.'
  - Relation: BORN\_IN(Marie Curie, Poland) → [PERSON] → [LOCATION]
- **Relation Extraction Methods:**
  - Pattern-Based: 'X was born in Y' → BORN\_IN(X,Y)
  - Supervised: Train classifier on labeled (entity1, relation, entity2) triples
  - Distant Supervision: Use knowledge base (Freebase, Wikidata) for auto-labeling
  - Neural: CNN/LSTM/BERT-based models with entity pair as input

*Note: Entity Extraction finds WHAT the entities are. Relation Extraction finds HOW they relate to each other. Together they populate knowledge graphs.*

#### Q4(c) — Coreference Resolution

[4 Marks]

- **Definition:** Coreference resolution identifies when different words/phrases in text refer to the same real-world entity.
- **Examples:**
  - 'Barack Obama visited India. He met the Prime Minister.' → He = Barack Obama
  - 'Apple released a new iPhone. The company is doing well.' → The company = Apple
  - 'John gave Mary the book. She read it immediately.' → She = Mary, it = the book
- **Key Terms:**
  - Mention: Any noun phrase that refers to an entity
  - Antecedent: The first/main mention of an entity
  - Coreference Chain: All mentions referring to the same entity
- **Methods:**
  - Rule-Based: Hobbs algorithm, gender/number agreement
  - ML-Based: Mention-pair models, entity-mention models
  - Neural: End-to-end neural coreference (SpanBERT, Lee et al.)

*Note: Coreference resolution is essential for information extraction, question answering, and dialogue systems to track entities across sentences.*

### November-December 2023

#### Q3(a) — Information Retrieval and NLP's Role

[4 Marks]

**[REPEATED] — Refer to: May-June 2023 → Q3(a) [IR system description]**

- **Additional — Significance of NLP in IR:**
  - Without NLP: keyword matching only (brittle, misses semantics)
  - With NLP: understands query intent, handles synonyms, resolves ambiguity
  - Example: Query 'car' should also retrieve documents about 'automobile', 'vehicle'

#### Q3(b) — Reference Resolution and Coreference Resolution

[8 Marks]

- **Reference Resolution (general term):** Process of determining what each referring expression in text refers to.
- **Types of Reference:**
  - Anaphora: Refers back to a previously mentioned entity ('she', 'it', 'they')
  - Cataphora: Refers forward to entity mentioned later ('Before he left, John packed.')
  - Bridging: Implicit reference ('I bought a book. The author was famous.' — author of book)

- **Coreference Resolution (specific case):**

**[REPEATED] — Refer to: May-June 2023 → Q4(c) [Coreference Resolution]**

- **Additional — Algorithm Steps:**
  - 1. Detect all mentions (noun phrases, pronouns)
  - 2. Compute compatibility (gender, number, semantic type)
  - 3. Cluster compatible mentions into coreference chains
  - 4. Resolve each chain to its canonical entity

*Note: Reference resolution is broader than coreference. Coreference resolution is specifically about different expressions referring to the same entity.*

**Q3(c) — Cross-Lingual Information Retrieval [6 Marks]**

**[REPEATED] — Refer to: May-June 2023 → Q3(c) [CLIR definition, approaches, example]**

**Q4(a) — Vector Space Model [6 Marks]**

**[REPEATED] — Refer to: May-June 2023 → Q4(a) [VSM full explanation with cosine similarity]**

**Q4(b) — Entity Extraction and Relation Extraction [8 Marks]**

**[REPEATED] — Refer to: May-June 2023 → Q4(b) [Entity extraction and relation extraction with examples]**

**Q4(c) — Named Entity Recognition (NER) [4 Marks]**

**[REPEATED] — Refer to: May-June 2023 → Q3(b) [NER definition, entity types, evaluation metrics]**

## May-June 2024

---

**Q3(a) — Vector Space Model (VSM): Strengths and Weaknesses [9 Marks]**

**[REPEATED] — Refer to: May-June 2023 → Q4(a) [VSM basics]**

- **Extended — VSM Document/Query Representation:**
  - Vocabulary: All unique terms across all documents
  - Document vector:  $[tf-idf(t_1, d), tf-idf(t_2, d), \dots, tf-idf(t_N, d)]$
  - Query vector:  $[tf-idf(t_1, q), 0, tf-idf(t_3, q), \dots, 0]$  for query terms
- **Detailed Strengths and Weaknesses:**

Strengths	Weaknesses
Simple, efficient, scalable	Ignores word order and grammar
Partial matching supported	No semantic understanding
Ranked retrieval (cosine score)	Synonym problem: 'car' ≠ 'automobile'
TF-IDF handles term importance	Very high-dimensional sparse vectors
Works well for keyword search	Cannot handle phrasal queries well

**Q3(b) — Evaluating NER Systems****[9 Marks]****[REPEATED] — Refer to: May-June 2023 → Q3(b) [NER metrics: Precision, Recall, F1]**

- **Additional — Detailed Evaluation Analysis:**
- **Entity-Level Evaluation (Exact Match):**
  - Both entity span and entity type must match exactly
  - Example: Predicted [New York | LOC], Actual [New York | LOC] → TP
  - Example: Predicted [New | LOC], Actual [New York | LOC] → FP and FN
- **Token-Level Evaluation:**
  - Each token's IOB tag is compared independently
  - More lenient than entity-level
- **How to Improve NER System (from result analysis):**
  - Low Precision: Model tags too many entities → add negative examples, tighten rules
  - Low Recall: Misses many entities → expand training data, add more entity types
  - Domain-specific errors: Fine-tune on domain corpus (medical, legal, etc.)
  - Boundary errors: Use CRF layer to enforce valid IOB transitions

**Q4(a) — Cross-Lingual Information Retrieval (CLIR): Challenges****[9 Marks]****[REPEATED] — Refer to: May-June 2023 → Q3(c) [CLIR basics]**

- **Extended — Machine Translation in CLIR:**
  - MT helps bridge the language gap by translating queries or documents
  - Query Translation (preferred): Cheap — translate only the query
  - Document Translation (expensive): Translate entire corpus
  - Cross-lingual embeddings: LASER, LaBSE — map all languages to shared space without explicit MT
- **Challenges Detailed:**

Challenge	Explanation	Solution
Translation ambiguity	Word has multiple meanings	Use context-aware MT (neural MT)
Morphological variation	Word forms differ (run/ran/running)	Lemmatization, stemming
OOV words	Named entities not in dictionary	Transliteration, character models
Low-resource languages	Insufficient parallel data	Cross-lingual transfer learning
Grammar differences	SOV vs SVO word order	Neural MT handles this implicitly

**Q4(b) — Entity Extraction: Difference from NER****[9 Marks]****[REPEATED] — Refer to: May-June 2023 → Q4(b) [Entity extraction basics]**

- **Extended — Entity Extraction vs NER:**

Aspect	Entity Extraction	NER
Scope	Broader — includes any entity type	Predefined categories (PER, ORG, LOC)
Flexibility	Domain-specific patterns	Standard taxonomy
Example	Product names, medical codes, URLs	People, places, organizations
Output	Raw entity strings + metadata	Entity string + type label

- **Real-World Applications of Entity Extraction:**

- Healthcare: Extract drug names, dosages, side effects from clinical notes
- Finance: Extract stock symbols, company names, financial figures from reports
- Legal: Extract case numbers, dates, parties from legal documents
- E-commerce: Extract product names, prices, specifications from web pages

## May-June 2025

---

### Q3(a) — Reference Resolution and Coreference Resolution

[8 Marks]

[REPEATED] — Refer to: Nov-Dec 2023 → Q3(b) [Reference and coreference resolution with examples]

- **Additional — How these help understand entity relationships:**
  - Reference Resolution maps all mentions to entities, creating a knowledge map of the text
  - Coreference resolution chains allow tracking an entity across paragraph boundaries
  - Essential for multi-document question answering and summarization
- **Extended Examples:**
  - 'The president signed the bill. He approved it despite opposition.' → He=president, it=bill
  - 'Amazon launched Alexa. The virtual assistant understands voice commands.' → The virtual assistant = Alexa

### Q3(b) — Cross-Lingual Information Retrieval (CLIR)

[8 Marks]

[REPEATED] — Refer to: May-June 2023 → Q3(c) [CLIR with additional from 2024 MJ Q4(a)]

### Q3(c) — What is Information Retrieval?

[2 Marks]

- **IR:** Finding relevant material (documents, web pages) from a large collection based on an information need (query).
  - Goal: Return the most relevant documents ranked by similarity to query
  - Used in: Search engines, digital libraries, email search, recommendation systems

**Q4(a) — Entity Extraction in Information Retrieval****[8 Marks]****[REPEATED] — Refer to: May-June 2023 → Q4(b) [Entity extraction] + May-June 2024 → Q4(b) [applications]**

- **Additional — Entity Extraction Techniques and Algorithms:**
- **Rule-Based:**
  - Regular Expressions: Pattern matching (e.g., dates: \d{2}/\d{2}/\d{4})
  - Gazetteers: Lookup lists of known entity names
  - Grammar rules: Capitalize proper nouns → entity candidate
- **ML-Based:**
  - Conditional Random Field (CRF): Sequence labeling model, considers context
  - Hidden Markov Model (HMM): Probabilistic sequence model
- **Deep Learning:**
  - BiLSTM-CRF: Bidirectional LSTM + CRF layer for IOB tagging
  - BERT-NER: Fine-tuned BERT with token classification head

*Note: Entity Extraction serves as the foundation for building knowledge graphs, search indexes, and intelligent information systems.*

**Q4(b) — Vector Space Model (VSM)****[8 Marks]****[REPEATED] — Refer to: May-June 2023 → Q4(a) [VSM full explanation]****Q4(c) — What is Named Entity Recognition (NER)?****[2 Marks]**

- **NER:** A subtask of information extraction that identifies and classifies named entities in text into predefined categories such as Person, Organization, Location, Date, etc.
  - Example: 'Elon Musk [PER] founded Tesla [ORG] in California [LOC]'

**November-December 2025**

---

**Q3(a) — Vector Space Model in Information Retrieval****[9 Marks]****[REPEATED] — Refer to: May-June 2023 → Q4(a) + May-June 2024 → Q3(a) [VSM full explanation]**

- **Extended — Relevance Computation Steps:**
  - 1. Index: Build inverted index mapping terms to documents
  - 2. Represent: Convert doc and query to TF-IDF vectors
  - 3. Compute: Cosine similarity between query vector and each document vector
  - 4. Rank: Sort documents by descending similarity score
  - 5. Return: Top-k documents to user



**Q3(b) — Entity Extraction vs Relation Extraction vs Coreference Resolution [8 Marks]**

- **Comparison Table:**

Task	What it does	Example	Output
Entity Extraction	Finds named entities in text	'Google' in news article	[ORG: Google]
Relation Extraction	Finds relationships between entities	'Google acquired YouTube'	ACQUIRED(Google, YouTube)
Coreference Resolution	Links different mentions of same entity	'Google...the company...'	Merge into one entity chain

- **How they contribute to building knowledge from text:**
  - Entity Extraction: Identifies the key actors/objects (nodes in knowledge graph)
  - Relation Extraction: Identifies connections (edges in knowledge graph)
  - Coreference Resolution: Merges duplicate nodes referring to same entity
- **Combined Example:**
  - Text: 'Sundar Pichai leads Google. He joined the company in 2004.'
  - Entity: [PER: Sundar Pichai], [ORG: Google]
  - Relation: LEADS(Sundar Pichai, Google), JOINED\_IN(Sundar Pichai, Google, 2004)
  - Coreference: He = Sundar Pichai, the company = Google

*Note: These three tasks together enable automatic construction of structured knowledge bases from raw unstructured text, used in search engines, QA systems, and recommendation systems.*

**Q4(a) — NER System Building Process****[9 Marks]****[REPEATED] — Refer to: May-June 2023 → Q3(b) [NER metrics and types]**

- **Full NER System Building — Supervised Approach:**
- **Step 1 — Data Collection:**
  - Collect domain-specific text (news, medical reports, legal docs)
  - Annotate with entity labels using tools like BRAT, Prodigy
- **Step 2 — Preprocessing:**
  - Tokenization → POS tagging → Sentence segmentation
- **Step 3 — Feature Engineering (for traditional ML):**
  - Word-level features: token text, POS tag, capitalization, suffix/prefix
  - Context features: previous/next word and its features
  - Gazetteer features: is word in known entity list?
- **Step 4 — Model Training:**

- CRF: Discriminative sequence model, uses transition + emission features
  - BiLSTM-CRF: Deep learning alternative, learns features automatically
  - BERT-NER: Fine-tune pre-trained BERT on labeled NER data
- **Step 5 — Evaluation:**
    - Use held-out test set → compute P, R, F1 per entity type
  - **Step 6 — Error Analysis and Improvement:**
    - Add more training data for error-prone categories
    - Use active learning to select most informative examples

*Note: Sample sentence: 'Tim Cook announced iPhone 15 in San Francisco last September.' → [PER: Tim Cook], [PRODUCT: iPhone 15], [LOC: San Francisco], [DATE: last September]*

#### Q4(b) — CLIR: Challenges and Approaches

[8 Marks]

[REPEATED] — Refer to: May-June 2023 → Q3(c) + May-June 2024 → Q4(a) [CLIR detailed]

### Topic Frequency Analysis — Unit 4

Topics ranked by how often they appear across all exam sessions (2023–2025):

Rank	Topic	Frequency	Sessions Asked
1	Cross-Lingual Information Retrieval (CLIR)	5×	MJ23, ND23, MJ24, MJ25, ND25
2	Vector Space Model (VSM)	5×	MJ23, ND23, MJ24, MJ25, ND25
3	Entity Extraction & Relation Extraction	5×	MJ23, ND23, MJ24, MJ25, ND25
4	Named Entity Recognition (NER) + Metrics	4×	MJ23, ND23, MJ24, MJ25
5	Coreference / Reference Resolution	4×	MJ23, ND23, MJ25, ND25
6	Information Retrieval (Introduction)	3×	MJ23, ND23, MJ25

*Note: ALL topics in Unit 4 are extremely high priority — every topic appears in 3-5 exam sessions. Focus on: CLIR (challenges + approaches), VSM (cosine similarity formula), NER (IOB tagging + P/R/F1), and Coreference Resolution (examples).*