NLP and Information Retrieval
Sample Exam 2

Generated by ChatGPT

Instructions

- Answer all questions.
- Show all calculations where required.
- Justify your answers clearly in transfer questions.

Task 1: Fundamentals of NLP

1a) (Knowledge)
Define the following terms in one or two sentences each:

- **Token**
- **Type**
- **Vocabulary**

1b) (Understanding)
Why might **linguistic preprocessing** (e.g., lowercasing, stopword removal) sometimes **harm** performance in deep learning pipelines, even though it helps in classical IR?

Task 2: Comparison of IR Models

2a) (Knowledge)
Contrast the **vector space model** (TF-IDF) and **probabilistic model** (BM25). How does BM25 address some shortcomings of a raw TF-IDF approach?

2b) (Transfer)
You run a small search engine for **multilingual** documents. Would you rely solely on BM25, or incorporate a semantic embedding-based technique? Justify your decision.

Task 3: True/False – Transformer & RNN

Mark each statement **True (T)** or **False (F)** and provide **one sentence of explanation**:

1. RNN-based LMs cannot model long-range dependencies at all.
2. Transformers require less memory than RNNs for the same sequence length.
3. Self-attention attends to all positions in the sequence simultaneously.
4. A GRU has more parameters than an LSTM for the same hidden size.
5. Transformers rely on positional encodings to track word order.

Task 4: N-Gram Smoothing

4a) (Knowledge)
What problem does **smoothing** solve in an n-gram language model, and why is it crucial for robust probability estimates?

4b) (Short Calculation)
Given a **bigram LM**, you have the counts:

- C(the, cat) = 10
- C(the, dog) = 0
- C(the) = 25

Apply **add-1 (Laplace) smoothing** to estimate $P(\text{dog} \mid \text{the})$. Show your steps.

Task 5: Byte-Pair Encoding (BPE)

5a) (Understanding)
Explain how **Byte-Pair Encoding** (BPE) is built step-by-step. Why does it help reduce **out-of-vocabulary** issues?

5b) (Mini Exercise)
Given the symbols {l, o, v, e, _} (underscore represents space), process the string "love love l ove" with **one or two hypothetical merges**. Show the updated tokenization.

Task 6: Retrieval Evaluation – MRR & nDCG

6a) (Calculation)
A ranked list of 5 documents has the following relevance:

- D1: relevant
- D2: relevant
- D3: not relevant
- D4: relevant
- D5: not relevant

Compute the **Mean Reciprocal Rank (MRR)**.

6b) (Understanding)
How does **nDCG** differ from MRR? When is nDCG preferable?

Task 7: Dense Retrieval & Approximate Nearest Neighbor

7a) (Knowledge)
Describe briefly how **ANN structures** like HNSW or Faiss speed up similarity searches in high-dimensional embeddings.

7b) (Transfer)
You plan to handle **100 million** paragraphs in a news archive. Which **ANN structure** (HNSW, IVF in Faiss, etc.) would you pick and why?

Task 8: Neural Re-Ranking with BERT

8a) (Understanding)
In a **two-stage retrieval** pipeline, how are query and document tokens typically combined as input for BERT?

8b) (Transfer)
Your e-commerce store has high user traffic. If you add a BERT re-ranker to rank the top 100 results, latency becomes a problem. Suggest **two optimizations**.

Task 9: Instruction Tuning

9a) (Knowledge)
What is **instruction tuning**? How does it differ from using long prompts in an **unmodified** model?

9b) (Transfer)
Your **FAQ chatbot** must follow formatting instructions. Describe **two ways** instruction tuning improves consistency, and **one limitation**.

Task 10: RLHF & Model Alignment

10a) (Understanding)
In **RLHF**, what is the role of the **reward model**, and why use **pairwise human feedback** instead of direct numeric scores?

10b) (Transfer)
Your chatbot produces **toxic** outputs in rare cases. How would you build a **human preference dataset** to correct this behavior? Mention one risk.

Task 11: Handling Long Context in Transformers

11a) (Knowledge)
What problem arises in **self-attention** when handling long sequences?

11b) (Transfer)
You need to answer questions about **25,000-token transcripts**. Would you use **sparse-attention Transformers** or **retrieval-augmented generation (RAG)**? Justify.