**Birla Institute of Technology & Science, Pilani**
**Work Integrated Learning Programmes Division**
**First Semester 2023-2024**
**M.Tech. in AIML**

**End-Semester Test**
**(EC-2 Regular Paper)**

| | | |
|---|---|---|
| Course No. | : AIML* ZG519 | |
| Course Title | : Natural Language Processing Applications | |
| Nature of Exam | : Open Book | No. of Pages = 2 |
| Weightage | : 30% | No. of Questions = 3 |
| Duration | : 2 Hours | |
| Date of Exam | : 21 April, 2024 | |

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

**Question 1. [11 Marks]**

**A. Statistical machine translation - 3 marks**
IBM Model 1 alignment - 3 marks

**Scenario:** We have a parallel corpus containing a simple sentence pair:

Source (English): I like coffee.
Target (Spanish): Me gusta el café.

Q1. How many possible alignments are there for this sentence pair, considering each English word can be aligned to a Spanish word or the null token (Ø)? [1.5]

**Solution:** There are three English words and four possible alignments for each (aligned to "Me", "gusta", "el", "café", or Ø). Therefore, the total number of alignments is: 3^(words in English) = 3^3 = 27

Q2. Imagine we have a uniform initial probability distribution for all alignments. Calculate the expected count (E-count) for the following alignment: [1.5]

I -> Ø, like -> gusta, coffee -> café

**Solution:** Since the initial probabilities are uniform (1/total alignments), the E-count for any specific alignment is:

E-count = (Number of sentences in corpus) / (Total number of alignments)

E-count = (Number of sentences in corpus) / 27 (Fill in the number of sentences in your corpus)

**B. Neural machine translation - 4 marks**

We have an NMT model trained on an English-French corpus. We are translating the English sentence "The cat sat on the mat." and are currently at the third word prediction stage (following "The" and "cat"). The model outputs the following probabilities for the next French word:

- "est" (is): 0.3
- "sur" (on): 0.4
- "le" (the): 0.2
- "un" (a/an): 0.1

Assume, the beam search width is set to 2 (we keep track of the top 2 most likely partial translations). Considering the current translation state and beam width, calculate the scores for all possible continuations (2-word sequences) and identify the top 2 scoring sequences.

**Solution:**

**Step 1: Calculate Scores for Continuations**

For each candidate French word, multiply its probability with the score of the best existing partial translation (since beam width is 2, we'll have two scores).

**Existing Partial Translation 1:** "The cat" (assume its score is S1)

- "The cat est": 0.3 * S1
- "The cat sur": 0.4 * S1

**Existing Partial Translation 2:** "The cat" (assume its score is S2, potentially different from S1)

- "The cat le": 0.2 * S2
- "The cat un": 0.1 * S2

**Step 2: Identify Top 2 Scoring Sequences**

The top 2 scoring sequences (considering both existing partial translations) will be the ones kept for the next beam search step.

**Example:**

- If S1 is higher than S2, and "The cat sur" has the highest score among all continuations, then the top 2 sequences could be:
  - "The cat sur" with score 0.4 * S1
  - "The cat est" with score 0.3 * S1 (assuming it's the second-highest score)

**C. a) Indic Machine translation- 4 marks**

how to implement machine translation for low resource indic languages like Assamese - 2 marks

Ans: Back-Translation (BT):
(1) Train a NMT model using a pre-trained model for target languages
(2) Translate monolingual English sentences into the target Indian language.
(3) filter and retain only high-quality ones.
(4) Add these high-quality translated sentence pairs to the original parallel corpus, effectively increasing training data.
(5)Multilingual Data Augmentation to use parallel corpora available for other Indian languages with more resources

b) Explain how transfer learning does work better for related languages like tamil? 2 marks

Ans: Transfer learning works better due to Shared Linguistic Features:
1)Family Similarities on vocabulary and grammatical structures.
2)Script Similarities
3)Cognitive Bias to connect to previously learned languages

4)Improved Feature Representation
5) Domain Adaptation: domain adaptation techniques might be necessary to bridge the gap and ensure accurate translations.

**Question 2. [11 Marks]**

**A NER- Do IE can assist in Knowledge graph OR do knowledge Graph assist in IE? Explain the argument with example - 1+1 marks**

Question 1 :Discuss the symbiotic relationship between Information Extraction (IE) and Knowledge Graphs (KGs). Can IE assist in the creation and enrichment of Knowledge Graphs, or does the existence of Knowledge Graphs enhance the performance of Information Extraction systems? Provide arguments for both perspectives and support your answer with examples. [3 Marks]

Ans: Information Extraction (IE) and Knowledge Graphs (KGs) have a mutually beneficial relationship, each contributing to the enhancement and refinement of the other.

IE assisting in Knowledge Graphs:
IE plays a crucial role in the creation and enrichment of Knowledge Graphs by extracting structured information from unstructured text data. By identifying entities, relations, and events from text, IE systems contribute valuable data points that can be incorporated into a Knowledge Graph. For example, consider an IE system extracting information about companies, their CEOs, and the acquisition events mentioned in news articles. These extracted entities and relations can be used to populate a Knowledge Graph representing the corporate world, thereby enriching its content and improving its coverage.

Knowledge Graph assisting in IE:
On the other hand, Knowledge Graphs serve as valuable resources for enhancing the performance of IE systems. By providing structured knowledge about entities, relations, and their connections, Knowledge Graphs offer valuable context and constraints that can guide the information extraction process. For instance, an IE system extracting information about movies and actors can leverage a movie-centric Knowledge Graph to validate extracted facts against existing knowledge and ensure their accuracy. Additionally, Knowledge Graphs can be used to resolve ambiguity and disambiguate entities mentioned in text, thereby improving the precision and recall of IE systems.

Overall, the integration of Information Extraction and Knowledge Graphs enables the creation of comprehensive, structured knowledge bases that capture diverse aspects of the world. This synergy between IE and KGs facilitates various applications, including semantic search, question answering, and knowledge discovery.

Alternate Question 1: Given the following sentences:

"Apple Inc. is headquartered in Cupertino, California."
"Tim Cook is the CEO of Apple Inc."
"Apple Inc. announced the launch of its new iPhone model."
"The event will take place at the Steve Jobs Theater."
Extract the named entities, relations, and events mentioned in these sentences. Provide a brief description of each named entity, relation, and event extracted from the sentences. [3 Marks]

Ans: Extracted Information:

Named Entities:

"Apple Inc.": Organization
"Cupertino, California": Location
"Tim Cook": Person
"CEO": Title
"iPhone": Product
"Steve Jobs Theater": Location
Relations:

(Tim Cook, CEO, Apple Inc.): Person (Tim Cook) holds the position of CEO in the Organization (Apple Inc.).
(Apple Inc., launch, iPhone): Organization (Apple Inc.) announced the launch of the Product (iPhone).
(event, location, Steve Jobs Theater): The event will take place at the Location (Steve Jobs Theater).
Events:

Apple Inc. Headquarters Establishment: Apple Inc. is headquartered in Cupertino, California.
CEO Appointment: Tim Cook is the CEO of Apple Inc.
Product Launch: Apple Inc. announced the launch of its new iPhone model.
Event Venue Booking: The event will take place at the Steve Jobs Theater.

Question 2: Consider an ACE model trained for Named Entity Recognition (NER) tasks on text data. The model undergoes evaluation on a test set consisting of 100 instances. Out of these instances, the model correctly identifies 70 instances as Named Entities (NEs), incorrectly identifies 10 instances as NEs, and misses 20 instances that are actually NEs. Calculate the precision, recall, F1-score, and accuracy for this ACE model. Provide your calculations and interpretations.[ 4 Marks]
Ans: To calculate the evaluation metrics for the ACE model, we first need to define the terms:

True Positives (TP): Instances correctly identified as NEs by the model (70).
False Positives (FP): Instances incorrectly identified as NEs by the model (10).
False Negatives (FN): Instances missed by the model that are actually NEs (20).
Total Instances (N): Total number of instances in the test set (100).
Precision= 0.875
Recall=0.778
F1−score =0.858
Accuracy=0.70

Interpretation:

Precision (0.875) indicates that among all instances predicted as NEs by the model, 87.5% are actually NEs. This suggests that the model is relatively precise in identifying NEs.
Recall (0.778) signifies that the model captures 77.8% of all actual NEs present in the test set. This indicates the effectiveness of the model in identifying NEs comprehensively.
F1-score (0.858) provides a harmonic mean of precision and recall, offering a balanced measure of the model's performance in terms of both precision and recall.
Accuracy (0.70) represents the overall correctness of the model's predictions, indicating that 70% of all instances in the test set are correctly classified.

Alternate question 2: Explain how the Viterbi algorithm is utilized to decode the most probable sequence of labels for the provided input sentence "The quick brown fox jumps over the lazy dog" in a Maximum Entropy Markov Model (MEMM) trained for part-of-speech tagging. [4 Marks]

Ans: The Viterbi algorithm is a dynamic programming technique used to find the most probable sequence of labels (or states) in a sequence labeling task, such as part-of-speech tagging. In the case of a Maximum Entropy Markov Model (MEMM), the Viterbi algorithm is employed to decode the sequence of part-of-speech tags that best matches the observed input sentence.

Here's how the Viterbi algorithm is utilized for the given input sentence "The quick brown fox jumps over the lazy dog" in a MEMM trained for part-of-speech tagging:

Initialization: For each word in the input sentence, the MEMM calculates the probability of each possible part-of-speech tag being assigned to that word. These probabilities are based on the features extracted from the word and its context.

Forward Pass: Starting from the beginning of the sentence, the algorithm iterates through each word in the input sequence. For each word, it calculates the probability of transitioning from the previous state (part-of-speech tag) to each possible current state (part-of-speech tag) using the MEMM's transition probabilities.

Backtracking: As the algorithm progresses, it keeps track of the most probable path leading to each state (part-of-speech tag) at each position in the input sequence. This information is stored in a dynamic programming table.

Decoding: Once the algorithm reaches the end of the input sequence, it identifies the most probable sequence of part-of-speech tags by backtracking from the final state to the initial state, selecting the path with the highest probability at each step.

By applying the Viterbi algorithm to the MEMM, we can efficiently determine the most likely sequence of part-of-speech tags for the provided input sentence. This allows us to perform accurate part-of-speech tagging, which is essential for various natural language processing tasks.

This process ensures that the MEMM selects the sequence of part-of-speech tags that maximizes the overall probability of the observed input sentence given the model's parameters and training data.

extra Question: Demonstrate how IOB encoding is applied to annotate the entities in the following sentence : "John lives in New York City." - 2 mark

Ans: "John" - B-PER (Beginning of a Person entity)
"lives" - O (Outside any named entity)
"in" - O (Outside any named entity)
"New" - B-LOC (Beginning of a Location entity)
"York" - I-LOC (Inside a Location entity)
"City" - I-LOC (Inside a Location entity)
"." - O (Outside any named entity)
So, the IOB-encoded version of the sentence would be:
"John B-PER lives O in O New B-LOC York I-LOC City I-LOC . O"

In this encoding:

"B-PER" denotes the beginning of a Person entity.
"B-LOC" denotes the beginning of a Location entity.
"I-LOC" denotes that the token is inside a Location entity.
"O" denotes that the token is outside any named entity.
This IOB-encoded representation provides a structured way to represent named entities in text data, facilitating the training and evaluation of NER models.

**Question 3. [ 8 Marks]**

A. A data scientist working for a social media analytics company needs to implement various techniques for sentiment analysis to ensure accurate and efficient analysis of vast amounts of text data collected from public opinion on various topics through social media platforms.
Describe two techniques that he can consider to use. Explain how each one works. Also, discuss the advantages and disadvantages of each technique in the context of processing large volumes of social media text data. (4 Marks)

Solution :

Following two techniques should be explained with advantages and disadvantages in the context of social media text data

- Rule-based sentiment analysis — 2Marks
- Machine learning based sentiment analysis — 2 Marks

B. Explain the concept of Target-dependent LSTM (TD-LSTM) and its application in capturing aspect information within sentences. Discuss the advantages of employing TD-LSTM over traditional LSTM architectures in aspect-based sentiment analysis tasks, and provide examples of real-world applications. (4 Marks)

Solution :
Explaining the concept of TD-LSTM and its application - 2 Marks
Advantages with real-world application - 2 Marks