Roll No												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

Department of Research & Development

Mid - Term Examinations - AUGUST 2024

Odd Semester: Ph.D. Course Work

Course Code: CSE863

Course Name: Advanced Natural Language Processing for Educational Applications

Department: Computer Science and Engineering

Date: 12/08/2024

Time: 09:30am – 11:00am

Max Marks: 50

Weightage: 25%

Instructions:

(i) Read the all questions carefully and answer accordingly.

(ii) Do not write any matter on the question paper other than roll number.

PART A (THOUGHT PROVOKING)

Answer all the Questions. Each question carries 5 marks. **(4Qx 5M= 20M)**

1. List out any 5 models of word embeddings / pre-trained language models. State whether they are static word embeddings or contextual word embeddings / pre-trained language models. NOTE: In your answer, you need to mention ***at least 1*** static word embedding and ***at least 1*** contextual word embedding / pre-trained language model. (CO:1 BL: Comprehension)

2. Mention any languages in which we currently have automatic essay grading datasets available. You can mention up to 10 such languages, but you will be getting full credit if any 5 are correct. (CO:1 BL: Comprehension)

3. About 10 years ago, we used ***static word embeddings***. In the last few years, we moved to ***contextual word embeddings*** and ***pre-trained language models*** (like BERT), and slowly, but surely to ***large language models*** (like GPT3.5 (a.k.a. ChatGPT)). Explain the main difference between the pre-trained language models (like BERT) and the earlier static word embeddings (like word2vec / GloVe), by selecting a ***polysemous word*** (word with multiple meanings) and mention ***at least 2 contexts*** which that word is used. (CO:2 BL: Comprehension)

4. Briefly explain how you would create a system to automatically grade essays. You need to briefly describe the system architecture, mention any necessary pre-processing steps (eg. Score normalization, lower-casing, etc.), evaluation metric used, etc. (CO:2 BL: Comprehension)

PART B (PROBLEM SOLVING)

Answer all the Questions. Each question carries 10 marks. **(3Qx 10M= 30M)**

5. A limerick is a poem which has 5 lines, with a rhyming scheme of AABBA (1st, 2nd and 5th line rhyme, as does the 3rd and 4th line rhyme separately). Prof. SAM decides to create a ***limerick classifier***, where, given a set of 5 lines, he has to decide whether or not they form a limerick. To do this, he

creates a ***static 26-dimension vector representation*** for ***the last word on each line***, such that each dimension stores the ***number of times that the character appears in the word***. For example, the vector for the word ***sam*** is given as:

Vector[“sam”] = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]

He then goes on to calculate the ***cosine similarities*** between each pair of ***last words*** for each pair of lines, using the formula:

$$CS(A,B) = \frac{\sum_{i=1}^{26} A_i * \ B_i}{\sqrt[2]{\left(\sum_{i=1}^{26} A_i^2\right) * \left(\sum_{i=1}^{26} B_i^2\right)}}$$

He concludes that the set of lines constitute a limerick if the ***cosine similarities*** of the ***last words*** of the lines (1,2), (1,5), (2,5), and (3,4) are the four highest cosine similarities of the 10 possible cosine similarity pairs. With that rule, decide whether or not the following set of 5 lines constitutes a limerick (convert all to lowercase while finding the vectors):

There was an old man from **Peru**,
Who dreamt he was eating his **shoe**,
In the middle of the **night**,
He woke up with a **fright**,
To see it was definitely **true**!

(CO:3 BL: Application)

6. Whistely et al. (2022) demonstrated a method of lexical simplification using a pre-trained language model (BERT), a part-of-speech tagger (NLTK) and fastText word embeddings. Explain how they used their system to simplify the word “**devoured**” in the sentence “The dog **devoured** its dinner.” You can assume that the top 5 candidates which are generated are “ate”, “wanted”, “was”, “eats”, and “had”. (CO:3 BL: Application)

7. Dutilleul et al. (2024) also described an approach to simplification using LLMs. Mention the LLM that they used. Explain the different prompting techniques that they used with an example. (CO: 3 BL: Application)