S20CS7.401:
Introduction to
Natural Language Processing

End Semester Exam

May 5, 2023

MM: 100 Time: 3 hrs

Note: Marks are mentioned next to the questions.

1. What are the advantages of LSTMs over RNNs? Clearly explain the architecture of LSTMs and GRUs.

   [15 marks]

2. In Attention based seq2seq translation methods, we see a tendency to repeat words or phrases in the output. What properties in the RNN derivative models might lead to this? How would you fix it?

   [15 marks]

3. Most Indo-Aryan and Dravidian (Indian) Languages exhibit rich morphology. This means, the words are attached with some grammatical markers. This results in a large number of word forms for words belonging to certain categories (noun, verbs (think of "eat" and all its forms in your native language). Typically, vector embedding methods take words as input. Can you design a vector embedding mechanism which can learn embeddings such that different morphological variations (for eg. walking, walks, walked, walk) of the same root (for eg. walk) are closer in the vector space?

   [15 marks]

4. Transliteration: The task is to transliterate words from a non-phonetic script (Roman. English) to a phonetic script (Indian scripts like Devanagari (*Ex.* Hindi) etc.) Please mention the challenges. Develop an NN model for the same. Clearly motivate your choice of model and also explain how the model will address various challenges (*Ex.* : First syllables of "Michael" vs. "Michelle")

   [20 marks]

5. Explain Multi-Headed Attention in Transformers in detail. Why do we need multiple heads?

   [20 marks]

6. Many words in a language have multiple meanings. Why can't Word2Vec, GloVe etc. provide the meaning representation for all meanings of a word form? What are the approaches that provide context dependent meaning representation? Explain one such approach in detail.

   [15 marks]