Instructions: Start with Part-A (closed book). As soon as you finish Part-A, you can take another answer sheet for Part-B (reference material as declared). There is no time limit for A and B. Manage your time on your own.
Give appropriate details in all the answers.

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2023-2024 CS F429 –Natural Language Processing
Comprehensive Exam Part-A (Closed Book)

Weightage: 22% **Date: 13th Dec 2023**

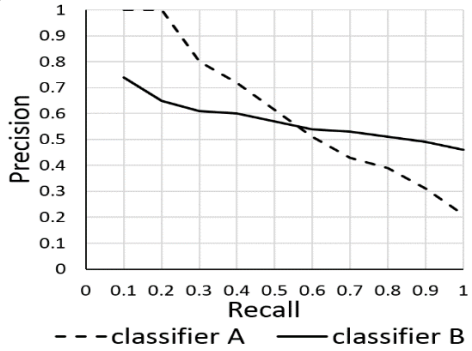
1. (a) If we update our word vectors when training the LSTM model on sentiment classification data, how would these word vectors differ from ones not updated during training? Explain with an example. Assume that the word vectors of the LSTM model were initialized using GloVe or word2vec.
   (b) How is the final word vector obtained from skip-gram? [2.5]
2. True/False with justification
   (a) Gradient clipping is an effective way of solving vanishing gradient problem.
   (b) Gated recurrent units (GRUs) have fewer parameters than LSTMs.
   (c) Neural window-based models can be parallelized, but RNN language models cannot.
   (d) Gradient of sigmoid activation function is increasing as the input input increases [4]
3. What is the role of QKV in attention network? What is the difference between self-attention and cross-cross attention? Explain. How does an attention network work differently than a recurrent network (LSTM)? [3]
4. Consider target-language sentence (English) *e* that corresponds to source-language (French) sentence *f*. Write a way to estimate word translation probability from e_i to f_j, $p(f_j/e_i)$ in IBM model 1, using parallel corpora. [2]
5. Compare BERT and T5 transformer models on the following with appropriate details:
   Architecture, attention, pretraining objectives, datasets, losses, fine-tuning, etc.
   Also discuss which one do you use for sentiment analysis and question answering system and why? [4]
6. Consider a two-bit register. The register has four possible states: 00, 01, 10 and 11. Initially, at time 0, the contents of the register is chosen at random to be one of these four states, each with equal probability. At each time step, beginning at time 1, the register is randomly manipulated as follows: with probability 1/2, the register is left unchanged; with probability 1/4, the two bits of the register are exchanged (e.g., 01 becomes 10); and with probability 1/4, the right bit is flipped (e.g., 01 becomes 00). After the register has been manipulated in this fashion, the left bit is observed. Suppose that on the first three time steps, we observe 0, 0, 1.
   Show how the register can be formulated as an HMM. What is the probability of transitioning from every state to every other state? What is the probability of observing each output (0 or 1) in each state? [2.5]
7. Suppose you are working on a farmer help line where queries posed by farmers are to be answered. Your job is to build a question-answer (QA) system. You have been given a pool of English paragraphs which may contain answer to some question(s). Draw a pipeline of the system with adequate details. [4]

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
I SEMESTER 2023-2024 CS F429 –Natural Language Processing
Comprehensive Exam Part-B (Open Book)

Weightage: 18% **Date: 13th Dec 2023**

1. Suppose that you are given the following sentences:
   - Chinese Beijing Chinese
   - Chinese Chinese Shanghai
   - Chinese Macao
   - Tokyo Japan Chinese

   (a) Learn a Bi-gram language model using this data with add-1smoothing.
   (b) Using the language model learnt in part (a) above, estimate the probability for the sentence, "Chinese Chinese Chinese Tokyo Japan". [4]
2. Word2Vec represents a family of embedding algorithms that are commonly used in a variety of contexts. Suppose in a recommender system for online shopping, we have information about co-purchase records for items x_1; x_2; : : : ; x_n (for example, item x_i is commonly bought together with item x_j). Explain how you would use ideas like Word2Vec to recommend similar items to users who have shown interest in any one of the items. [2]
3. (a) Consider the following interpolated precision-recall curves (see Figure) for two search engines that index research articles. Both are same search engines having different scoring methods for documents. If you want to have all relevant articles (not to miss even one) on some topic. Which engine would you prefer and why?
   (b) Find MAP for classifier A for a query (results given in the graph). What do you think what is important in getting high MAP? [3]



4. There are many sport events organized by various national and international agencies for various sports. Also, there are millions of *news items* on the different sport events or/and on the different sports by different agencies for last many years. Each news has short description of the sport(s) or/and the associated event details (if any), outcome of the event, etc. A news item can have one or more notes (reactions/feedback) associated with it and one or more sports associated with it. You are now asked to build a learning model which can infer the sport along with the particular event from the news item and associated notes. You have also been given the true sport and event associated with an item and note. Please note the following:
   - A note associated with the news items may not have proper sentences and there are 100 thousand sport and event combinations in total.
   - A news item may not have any event associated with it.

   Answer the following questions while building the model:
   (a) Which word representation would you use for your model and why?
   (b) How do you model the representation of the news item (along with its notes)?
   (c) Do you need to normalize your word vectors for a standard NN?
   (d) Which model would you use and why? How it is better than a standard feed forward NN?
   (e) Which output unit would you use and why?
   (f) How would you reduce the training and testing time of your model as you know that the total number of sport+event pair is very high? [9]