Name: _____ **Entry No**: _____

Indian Institute of Technology Jammu
Comprehensive Examination
PhD

Course Title: Natural Language Processing **Maximum Time: $2\frac{1}{2}$ Hrs**
Passing Marks: 30 **Maximum Marks: 60**

Instructions:

- Conditions of Examination: Closed Book/ Closed Notes; No discussion is allowed.

---

1. **[6 + 8 + 6 + 5 Marks]**

   TAG SET: NN – Noun of any form, MV – Main Verb, AV – Auxillary Verb, IN – Preposition, DT – Determiner, JJ – Adjective, CD – Number/Count

   TRAINING CORPUS:

   S1: Cricket/NN is/AV a/DT bat-and-ball/JJ game/NN played/MV between/IN two-teams/NN of/IN eleven-players/NN on/IN a/DT cricket-field/NN ./.

   S2: Baseball/NN is/AV a/DT bat-and-ball/JJ game/NN played/MV between/IN two-teams/NN of/IN nine-players/NN ./.

   S3: The/DT BJP-origin/NN lies/MV in/IN the/DT Bharatiya-Jana-Sangh/NN ./.

   S4: Congress/NN was/AV founded/MV in/IN 1885/CD during/IN the/DT British/JJ Raj/NN ./.

   S5: My-friend/NN in/IN India/NN brought/MV fruits/NN ./.

   TEST CORPUS: The British brought cricket in India in 1700s.

   a) [6 Marks] Develop the TF-IDF matrix and skip-gram representation for the above corpus for transforming to vector representation.

   b) [8 Marks] Design the Hidden Markov Network (HMM) for the given training data.

   c) [6 Marks] Apply Hidden Markov Model on the given training corpus, to compute the tag set for the given test corpus.

   [5 Marks] Compute the joint probability for a modified HMM where each hidden state is dependent on previous three states and the observed states depends only on its current state.

2. **[5 + 5 + 5 + 5 Marks]**

   You are given a problem of translating a sentence in language A to a sentence in language B. Design a learning model for this translation task using word alignment. In the word alignment, the model represents word alignment between a sentence in a foreign source language $A = a_1, a_2, \ldots, a_m$ and its target language $B = b_1, b_2, \ldots, b_n$. Here, m and n are the number of words in sentence of language A and language B respectively. Every source word a_i is aligned to one target word b_j and the alignments are are represented as a vector v (of the same length as the source) with $v_i = j$. Explain the train data and its format that your proposed model will use. Explain the complexity of your proposed model. Draw the computation graph for the proposed model. Also, suggest the utilization of Knowledge Graph (KG) for this alignment problem in the context of translation.

3. **[3 + 2 Marks]**

Given a seq2seq neural machine translation system with the following characteristics:
Source vocabulary size: 1000
Target vocabulary size: 2000
Encoder hidden state size: 100
Decoder hidden state size: 200
Attention mechanism is implemented by a one layer feedforward neural network.
RNN is implemented using an LSTM.
There is no embedding layer.
Assume no bias.

[Note:In the encoder, all four gates of LSTM are connected to input and previous states and in the decoder, all four gates are connected to encoder hidden state, previous state and previous output.]

(a) [3 Marks] Compute the sizes of all weight matrices involved in this seq2seq NMT model.

(b) [2 Marks] Discuss the importance of Positional Encoding in seq2seq models.

4. **[2 + 2 + 2 + 2 Marks]**

(a) [2 Marks] After watching training and validation loss function value at every epoch during training, how will you detect that your model is suffering from vanishing/exploding gradients or underfitting/ overfitting. Also, suggest an strategy to overcome these issues.

(b) [2 Marks] How do probabilistic CFGs (PCGFs) help with the ambiguity problem? What is the problem if some rule has a probability of 0.0? How is it overcome?

(c) [2 Marks] What will happen to learning if all the weights are initialized with a constant value c in a L-layered NN model? Suggest an appropriate weight initialization in the models for solving seq2seq problem.

(d) [2 Marks] Derive the formula to find the maximum number of continuous n-grams in a sentence with T tokens.

(e) [2 Marks] How many quad-grams phrases can be generated from the following sentence, after performing stop words and punctuation ($-$, . etc.) removal: Cricket is a bat-and-ball game played between two-teams of eleven-players on a cricket-field.