CS563 - NLP

ASSIGNMENT-2: Named Entity Recognition

(Read all the instructions carefully & adhere to them.)

Date: 12th Feb, 2021 **Deadline: 19th Feb, 2021**

Total Credit: 30

Instructions:

1. The assignment should be completed and uploaded by **19th Feb, 2021, 11:59 PM IST.**
2. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
3. Proper indentation and appropriate comments are mandatory.
4. You should zip all the required files and name the zip file as: **<roll_no>_assignment_<#>.zip, eg. 1501cs11_assignment_01.zip.**
5. Upload your assignment (**the zip file**) in the following link: https://www.dropbox.com/request/FWnXpXsRAiS6ZXHQhgXG

For any queries regarding this assignment you can contact:
Dushyant Chauhan (dushyantchauhan27@gmail.com) or
Kshitij Mishra (kmishra.kings@gmail.com)

ProblemStatement:

In most of the Information extraction (IE) pipelines, Named entity recognition (NER) is one of the first steps. It seeks to locate and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. For example: An unstructured block of text may be like:

Jim bought 300 shares of Acme Corp. in 2006.

After passing it through an NER system, the annotated block of text could be like:

$[Jim]_{Person}$ *bought 300 shares of* $[Acme\ Corp.]_{Organization}$ *in* $[2006]_{Time}$.

In this assignment, you have to develop different NER systems using Hidden Markov Model (HMM), Vanilla Recurrent Neural Network (RNN), Long short-term memory (LSTM) and Gated recurrent unit (GRU).

Setups:

1. Use the BIO tagging scheme to denote the beginning, intermediate and outside named entity.
2. Identify the named entities
3. Classify the named entity into the following types: person, product, company, geolocation, movie, music artist, tvshow, facility, sports team and other

Dataset:

- NER-Dataset-Train.txt: Training set
- NER-Dataset-10Types-Train.txt
- NER-Dataset-TestSet.txt: Test set
- Format:
  - Each line contains <Word \t Tag>
  - Sentences are separated by a blank line.
- Datasets can be downloaded from below given link: https://drive.google.com/drive/folders/1NYeUaJkhv5LpvUafTgYtZBErMyTka8aq?usp=sharing

Using the above mentioned dataset, perform the tasks mentioned in setups for the following four models:

1. **HMM based Model**
   a. **HMM Parameter Estimation**
      Input: Annotated tagged dataset
      Output: HMM parameters
      Procedure:
      Step1: Find states.
      Step2: Calculate Start probability (π).
      Step3: Calculate transition probability (A)
      Step4: Calculate emission probability (B)
   b. **Features for HMM**
      Use a bigram model
   c. **Testing**
      After calculating all these parameters apply these parameters using the Viterbi

algorithm, and determine the best sequence of named entity.

2. **Vanilla RNN based model**
    a. **Model Architecture**
    Draw a model architecture of the model you are proposing
    b. **Features for RNN**
    Please build features according to your understanding and choice
3. **LSTM and GRU based model**
    Same as for vanilla RNN

Evaluation (For all the models):

1. Perform 5 fold cross-validation on the Training datasets and report both average & individual fold results (Accuracy, Precision, Recall and F-Score). Also highlight and show the class-wise results of best-performing fold.
2. Submit Test Set Predictions (a total of 8 files, 2 for HMM, 2 for RNN, 2 for LSTM and 2 for GRU). [***have to upload***]
3. Write a report (doc or pdf format) on how you are solving the problems as well as all the results including model architecture (if any). [***have to upload***].