CL3.101 Computational Linguistics 1
Assignment 3 - Part 2

Course Instructor: Parameswari Krishnamurthy

Deadline: 6th April 2024, 9:00 PM

Instructions

- Your assignment must be implemented in Python. Do NOT use any standard library.
- Make sure the submitted assignment is your original work. Do not copy any part of the assignment from your friends. Do not refer any AI systems to generate the code.
- No deadline extension will be possible. Please start early in order to finish it on time.
- Make sure to follow the submission format properly. You will be penalised for not following the naming and submission format.

1 Introduction

For this part, you will be implementing and training CRF++ and HMM models for Part of Speech (POS) tagging. The training data for these models will be provided to you. Test the trained models on the annotated data from Part 1 and calculate the F1 scores for both models.

1.1 Training of Models

Train a CRF++ Model and an HMM Model for POS tagging task using the provided training data for both the languages. Here is the link to the training data `https://iiitaphyd-my.sharepoint.com/:f:/g/personal/sankalp_bahad_research_iiit_ac_in/EsxNff-0tUtJsSCUajLmt70Bcsvgrjsja6dLAkPVsxk_-Q?e=NPxB7h`. Incase your dataset does not have BIS tags, use a mapping from that tagset to BIS.

1.2 Theory

Explain the approaches used in CRF++ and HMM models for POS tagging. Provide a comparative study of these models along with your observations.

1.3 Testing and Analysis

Test both the CRF++ and HMM models on the annotated data from Part 1. Calculate the Precision, Recall and F1 scores, and give a confusion matrix for both models. Provide an analysis of the results.

2 Submission Format

Submit a single zip file named **RollNo_FirstName_POS_Models.zip** in which you will include the following files:

- Model files
- Analysis.txt (or .pdf) having the test set scores and observations
- ReadME.md containing instructions and details

Kindly adhere to this submission format properly.