Assignment 4: Parts-of-Speech Tagging

Course: Computational Linguistics - 1

Deadline: March 21st, 2025 — 23:59

1 General Instructions

1. The assignment must be implemented in Python. Do NOT use any standard libraries for HMM. You can use standard libraries for CRF.
2. Submitted assignment must be your original work. Please do not copy from any source.
3. Points distribution is provided for each section beforehand to avoid any confusion.
4. A single .zip file needs to be uploaded to the course portal.
5. Your grade will depend on correctness of implementation, and based on completion of all requirements specified in this document.

2 Introduction

In this assignment, you will explore Part-of-Speech (POS) tagging by implementing two widely used approaches: Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). Your task involves both annotating a dataset from scratch (which will later serve as your test data) and training a POS tagger using the provided training data. You will evaluate your models on the annotated test data and report results.

3 Part-1: Annotation

This section doesn't require any code. You are expected to do all the tasks manually.

3.1 Data Annotation

- You will manually annotate ∼500 tokens in **English** and ∼500 tokens in any **Indian language**. In total, you will have ∼1000 tokens annotated with tags.
- If you are not familiar with any Indian language, you may choose another non-English language.
- The annotation must follow the **BIS POS Tagset**. Detailed documentation can be accessed in reference section.
- Your dataset must be **original**—do not use existing POS-tagged datasets. Instead, collect data from sources such as **news articles**. Clearly mention the source of your chosen text in documentation.

3.2 Annotation Format

- Tokenize the sentences into words and store annotations in a text file.
- Each line should follow the format:

```
word \t tag
```

  i.e., Word and Tag separated by a tab.

- Example:

```
India    NNP
,        PUNC
Australia NNP
and      CC
England  NNP
are      VAUX
the      DT
Big      JJ
Three    CD
in       IN
Cricket  NN
.        PUNC
```

3.3 POS Tag Frequency Distribution

- Generate a **frequency distribution graph** of the POS tags for both languages.
- Ensure that the dataset is **as balanced as possible**, meaning the POS tags should be **evenly distributed** across the dataset.

4 Part-2: HMM and CRF Modeling

4.1 Training of Models

Train a **Hidden Markov Model (HMM)** model and a **CRF** for the POS tagging task using the provided training data for both languages.

- Training data can be accessed at: `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5787`
- In case your dataset does not use BIS tags, use a mapping from the existing tagset to the BIS tagset.

4.2 Theory

Explain the theoretical background of POS tagging approaches using **CRF** and **HMM**. Your explanation should cover:

- The fundamental principles behind each approach.
- The difference between **probabilistic models** (HMM) and **discriminative models** (CRF).
- The strengths and weaknesses of each method in the context of POS tagging.

Provide a **comparative study** of the two models along with your **observations** on their performance.

4.3 Testing and Analysis

Once the models are trained, test both the **HMM** and **CRF** models on your manually annotated test set (from Part 1). Perform the following evaluation:

- Calculate the **Precision, Recall, and F1-score** for both models.
- Generate a **Confusion Matrix** for both models.
- Provide a detailed **analysis of the results**, discussing:
  - Which model performed better and why?
  - The types of errors observed (e.g., confusion between noun and proper noun, auxiliary verbs misclassification, etc.).
  - Potential improvements to increase model accuracy.

5 Submission Guidelines

Submit a zip file named `<roll_number>_assignment4.zip` containing:

- Annotation of English Text (eng.txt)
- Annotation of Indian Language Text ({LANG}.txt)
- Graphs for Frequency Distribition (eng.png, {LANG}.png)
- Model checkpoints
- README.md with:
  - Documentation/Report containing Results and Analysis (either here or in Report.pdf)
  - Any assumptions or limitations

6 Resources

- BIS POS Tagset Documentation:
  `https://tdil-dc.in/tdildcMain/articles/134692Draft%20POS%20Tag%20standard.pdf`
- HMM - Interactive Illustration
  `https://nipunbatra.github.io/hmm/`
- Training dataset
  `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5787`