Assignment 5: Named Entity Recognition and Chunking

Course: Computational Linguistics - 1

Deadline: April 22nd, 2025 — 23:59

1 General Instructions

1. The assignment must be implemented in Python.
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, we will explore two important NLP tasks: Named Entity Recognition (NER) and Chunking. The assignment is divided into two parts: Part-1 covers NER and Part-2 covers Chunking. Since, POS-tagging and NER, Chunking are similar tasks, you can choose to modify your POS implementation for this assignment.

3 Part-1: Named Entity Recognition (50 points)

3.1 Annotation (15 points)

- Use the text assigned and annotated earlier during in-class NER annotation. You can choose to add more text from same wikipedia article.
- Use the following entity types: Person (PER), Organization (ORG), Location (LOC), and Miscellaneous (MISC).
- Ensure that annotation is accurate. For mapping the ILMT NER tagset to CoNLL tagset, use:

```
"NEP": "PER",
"NEL": "LOC",
"NEO": "ORG",
"NEDA": "MISC",
"NETI": "MISC",
"NEAR": "MISC",
"NEF": "LOC" or "ORG",  # depends on context
"NEU": None,           # no CoNLL equivalent
"NEMI": "MISC",
"NEN": None,           # numbers not annotated
"NETE": None           # dropped
```

- Use the BIO tagging scheme:
  - B-X: Beginning of entity type X
  - I-X: Inside (continuation) of entity type X
  - O: Outside any entity
- Example format:

```
Mahatma     B-PER
Gandhi      I-PER
visited     O
Ahmedabad   B-LOC
and         O
Delhi       B-LOC
during      O
his         O
journey     O
with        O
Indian      B-ORG
National    I-ORG
Congress    I-ORG
.           O
```

3.2 NER Model Implementation (25 points)

- Implement a sequence labeling model for Named Entity Recognition. (10 points)
- Use either Hidden Markov Model (HMM) or Conditional Random Field (CRF) (your choice).
- Train your model on the CoNLL-2003 English NER dataset (link provided in Resources). (5 points)
- Test your model on your manually annotated test set. (5 points)
- Report the following metrics: (5 points)
  - Precision, Recall, and F1-score
  - Confusion matrix for entity types

3.3 Error Analysis and Exploration (10 points)

- Analyse the results obtained. Conduct an error analysis identifying common error patterns (e.g., boundary errors, type confusion, etc.). Include this in README or Report (PDF). (7 points)
- Suggest potential improvements to address the observed errors. (3 points)

4 Part-2: Chunking (50 points)

4.1 Annotation (15 points)

- Manually annotate the same text used for NER, now for Chunking.
- Use the chunk types: Noun Chunks (NP), Verb Chunks (VGF, VGNF, VGINF, VGNN), Adjectival Chunks (JJP), etc. Refer to lecture slides for clarity.
- Use the IOB format:
  - B-X: Beginning of chunk type X
  - I-X: Inside (continuation) of chunk type X
  - O: Outside any chunk
- Example format:

```
Mahatma     B-NP
Gandhi      I-NP
visited     B-VGF
Ahmedabad   B-NP
and         B-CCP
Delhi       B-NP
during      B-NP
his         I-NP
journey     I-NP
with        B-NP
Indian      I-NP
National    I-NP
Congress    I-NP
.           O
```

4.2 Chunking Model Implementation (25 points)

- Implement a sequence labeling model for Named Entity Recognition. (10 points)
- Use either Hidden Markov Model (HMM) or Conditional Random Fields (CRF) (your choice).
- Train the model on the CoNLL-2000 Chunking dataset (link provided in Resources). (5 points)
- Test your model on your manually annotated test set. (5 points)

- Report the following metrics: (5 points)
  - Precision, Recall, and F1-score (both overall and per chunk type)
  - Confusion matrix for chunk types

4.3 Analysis of Chunking Patterns (10 points)

- Analyze which patterns or rules are most effective for different chunk types.
- Identify challenging chunking scenarios and discuss why they are difficult.
- Suggest potential improvements to address the observed errors.

5 Submission Guidelines

Submit a zip file named `<roll_number>_assignment5.zip` containing files in below preferred format:

- Annotation files:
  - NER annotations (ner_annotations.txt)
  - Chunking annotations (chunk_annotations.txt)
- Implementation code:
  - NER model (ner_model.py)
  - Chunking model (chunking_model.py)
  - Evaluation scripts (eval.py)
- README.md with:
  - Clear instructions for running your code
  - Explanation of model architecture
  - Data sources and preprocessing steps
- Report.pdf with:
  - Evaluation results
  - Error analysis
  - Conclusions and suggestions for improvement

6 Resources

- Download the datasets here:
  `https://drive.google.com/drive/folders/13waPgEKNxYrFPGOtr5fYtL38uOIkQoOJ?usp=sharing`
- CoNLL-2003 NER paper for reference:
  `https://aclanthology.org/W03-0419.pdf`
- CoNLL-2000 Chunking paper for reference:
  `https://aclanthology.org/W03-0419.pdf`