Assignment 2: Chunking

CS626: Speech and Natural Language Processing and the Web

Problem statement

Your goal is to implement a "chunker" in Python. You are supposed to use the 'conll2000' dataset provided with the assignment announcement. We have provided it in the ZIP file uploaded on Moodle/Teams. It is taken verbatim from the NLTK 'conll2000' dataset.

The sample data (one sentence) looks like:

```
The DT B-NP
Los NNP I-NP
Angeles NNP I-NP
Red NNP I-NP
Cross NNP I-NP
sent VBD B-VP
2,480 CD B-NP
cots NNS I-NP
, , O
500 CD B-NP
blankets NNS I-NP
, , O
and CC O
300 CD B-NP
pints NNS I-NP
of IN B-PP
Type-O JJ B-NP
blood NN I-NP
. . O
```

As a pre-processing step, you are supposed to remove the NP/VP/Other tags and only use B and I tags. The processed files are going to act as your input. **The expected output from the chunking task is divided into three parts:**

1) **Overall Precision, Recall, and F-score.**
2) **Precision, Recall, and F-scores for each tag (B and I).**
3) **A report on error analysis performed.**

The classifiers you need to implement are:

1) MEMM (You can use pre-implemented)
2) CRF (You can use pre-implemented)
3) BiLSTM (already implemented during the POS-tagging assignment, use that)

Note: You should understand the theory very well, though you are using the pre-implemented codes.

The feature set(s) which may be used are:

1) Current Word, Previous Word, Previous to Previous Word (word vectors).
2) Current POS, Previous POS, Previous to Previous POS.
3) Previous 2 chunk labels, previous to the current position.
4) Morphological features especially word stem and word affixes.

Please use the provided train/test splits for uniformity.

Dataset Statistics:
Train File (no. of lines) - 220663
Test File (no. of lines) - 49389

Additional Remarks:
No deadline; Evaluations as per the method proposed by the course instructor.
Just report progress at the second evaluation stage and onwards.

References:

- Shallow Parsing using Specialized HMMs
- Learning Better Internal Structure of Words for Sequence Labeling
- Bidirectional LSTM-CRF models for sequence tagging