Named Entity Recognition

Caleb L'Italien

Abstract

This project focuses on Named Entity Recognition (NER), which is the process of identifying and classifying named entities in text into predefined categories such as names, locations, and organizations. The project involves implementing two NER systems using two approaches: a classification-based approach and a sequence tagging approach. The models are evaluated using phrase-based scoring and the CoNLL evaluation script. The project also involves experimenting with different features to improve the performance of the models.

1 Introduction

Named Entity Recognition (NER) is a critical task in Natural Language Processing (NLP) that involves identifying and classifying named entities in text into pre-defined categories. NER is important for various NLP applications such as information extraction, question answering, and sentiment analysis.

The classification-based approach is already implemented in the starter code provided, which uses a logistic regression classifier from sklearn. My task is to experiment with different features that might help with NER and evaluate their performance using phrase-based scoring. The sequence tagging approach involves expanding the classifier so that it is trained to discriminate based on the probability of assigning a label to a word given its feature representation and the tag assigned to the previous word. I use the Viterbi algorithm to decode the tag sequence.

In summary, this project aims to develop two NER systems using two approaches and experiment with different features to improve their performance. The performance of the models are measured with accuracy, precision, recall, and f-scores.

2 NER as a Classification Task

The features tested on the model in the classification approach are: the word itself; the word existing in a name gazetteer provided by census.gov; the word containing an equals sign, hyphen, or apostrophe; the word being alphanumeric; the number of vowels in the word; the word being capitalized; and the word being entirely capitalized. The model performed on the development set with solely the word as a feature with an F-score of 54.00 percent. Each feature was then run individually (with the word as a feature as well) to see how they performed against this 54.00 percent benchmark. The features performed as follows: name gazetteer, +0.99; has an equals sign, +0.00; alphanumeric, -0.07; capitalized, +5.59; all capitalized, -0.98; has a hyphen, -0.80; has an apostrophe, -0.05; number of vowels, +4.61; and word length, -0.38. The model was then run on the test set with only features that produced a positive score (not including +0.00), as well as with all the features. The results of using only features producing a positive score on the development set are shown in Figure 1, and using all the features in Figure 2.

```
processed 51533 tokens with 3558 phrases; found: 4063 phrases; correct: 2459.
accuracy:  96.08%; precision:  60.52%; recall:  69.11%; FB1:  64.53
              LOC: precision:  68.63%; recall:  71.03%; FB1:  69.81  1122
             MISC: precision:  28.61%; recall:  29.20%; FB1:  28.91  346
              ORG: precision:  57.09%; recall:  70.79%; FB1:  63.20  1736
              PER: precision:  69.73%; recall:  81.50%; FB1:  75.16  859
```

Figure 1

```
processed 51533 tokens with 3558 phrases; found: 4034 phrases; correct: 2468.
accuracy:  96.18%; precision:  61.18%; recall:  69.36%; FB1:  65.02
              LOC: precision:  65.77%; recall:  69.83%; FB1:  67.74  1151
             MISC: precision:  29.56%; recall:  31.56%; FB1:  30.53  362
              ORG: precision:  58.74%; recall:  70.36%; FB1:  64.02  1677
              PER: precision:  73.34%; recall:  84.22%; FB1:  78.40  844
```

Figure 2

As can be seen in the figures, although certain

features did negatively impact performance on their own, performance improved using more features.

3 NER as a Sequence Tagging Task

The features tested on the model in the sequence tagging approach are the same as the ones tested on the classification approach. The model performed on the development set with solely the word as a feature with an F-score of 78.81 percent. This benchmark shows a clear improvement in performance through incorporating the previous words' label into the present words' features. Each feature was then again run individually (with the word as a feature as well) to see how they performed against this 78.81 percent benchmark. The features performed as follows: name gazetteer, +0.26; has equals, +0.00; alphanumeric, +1.32; capitalized, +3.44; all caps, -1.73; has hyphen, +0.26; has apostrophe, +0.00; number of vowels, -0.81; and word length, +2.11. The model was then run on the test set, again once with only features producing a positive score (not including +0.00), as well as with all the features. The results of using only features producing a positive score on the development set are shown in Figure 3, and using all the features in Figure 4.

```
processed 2294 tokens with 170 phrases; found: 167 phrases; correct: 139.
accuracy:  97.08%; precision:  83.23%; recall:  81.76%; FB1:  82.49
             LOC: precision:  83.02%; recall:  81.48%; FB1:  82.24  53
            MISC: precision:  80.00%; recall:  60.00%; FB1:  68.57  15
             ORG: precision:  85.07%; recall:  93.44%; FB1:  89.06  67
             PER: precision:  81.25%; recall:  74.29%; FB1:  77.61  32
```

Figure 3

```
processed 2294 tokens with 170 phrases; found: 167 phrases; correct: 136.
accuracy:  96.73%; precision:  81.44%; recall:  80.00%; FB1:  80.71
             LOC: precision:  78.18%; recall:  79.63%; FB1:  78.90  55
            MISC: precision:  84.62%; recall:  55.00%; FB1:  66.67  13
             ORG: precision:  84.62%; recall:  90.16%; FB1:  87.30  65
             PER: precision:  79.41%; recall:  77.14%; FB1:  78.26  34
```

Figure 4

As can be seen in the figures, using all features did not help performance, but only using features that individually produced positive impacts produced the highest performance.