CS388: Natural Language Processing Mini 1

Nidhi Kadkol
University of Texas at Austin UTEID: nk9368
nidhikadkol@gmail.com

1 Classification Method and Features

This report outlines the method used for Named Entity Recognition in which we focus on identifying tokens in a sentence which are person names. For this problem, I used logistic regression with Stochastic Gradient Descent (SGD). For the gradient update and score calculation, I made use of the SGDOptimizer class that was provided in the framework code. I used the following distinct features:

1. **Length:** Bucketized the lengths of the words into less than 5, between 6 and 10 inclusive, and so on. The individual lengths of the words are also included as separate features, as this increases the accuracy.
2. **Current Word:** The word itself is used as a feature.
3. **Word Appearance Count:** The number of times the word appears in the sentence (once, twice, or more than twice) helps to differentiate between names (used once or twice in a sentence) and other words, especially stop words like "of", "the", "and", etc. which appear often.
4. **Position at beginning and end:** 2 features called "first word" and "last word" which are true if the word is at the beginning or the end of the sentence respectively.
5. **Quotes:** If quotes are present in the sentence, it mostly means that someone is talking, and that person's name is likely to be there in the sentence.
6. **Summarized Pattern:** The words can be mapped to their structures in the form of patterns (David Nadeau, Satoshi Sekine, A survey of named entity recognition and classification, 2007). Capital letters are mapped to 'A', small letters to 'a', digits to '0' and then the patterns are condensed. For example, "Apple-Man" becomes "Aa-Aa", "12/07/1996" becomes "0/0/0".
7. **Prev and Next:** This is useful for getting context for the words, so we consider previous word, next word, previous and next word, and the same thing for patterns.
8. **New Word:** This feature is used for the test set. If a word has not been seen in the training set then when it is encountered in the test set this feature is true.

I used a learning rate of 0.1 and 58 epochs.
Collaborators - Abby Chua and Ankur Garg for learning rate and giving debugging tips.

2 Results

Table 1: Results

Set	F_1
Training	98.6%
Development	91.2%