Souvik Bagchi / Ikhlas Attarwala
05-24-19 / SLM
Project Status Report

A short recap:

Our project takes an approach to classify and **recognize named entities** within a large annotated corpus by means of a bidirectional LSTM. LSTMs are fantastic at tackling long-term dependencies, and while LSTMs preserve information of the past by using inputs read from left to right, bidirectional LSTMs use two hidden states to preserve information from both the past and future. Our dataset from Kaggle (click here for link) contains everything we need.

Statistical Summary:

The contents of the dataset we're using have been cleaned and annotated to reveal a word and its position in the corpus, along with two words both preceding and following it, where all 5 words' part-of-speech tag and shape (upper/lowercase, number, punctuation, etc.) have been provided. Additionally, each token has an IOB annotation for indicating the position within a chunk. The dataset contains 1,048,574 tokens using 35,179 unique words, split into 47,959 sentences, with 8 types of labels with 17 types of sub labeled named entities.

Results Summary and Evaluation Metrics:

Find our evaluation of the 8 labels so far.

	precision	recall	f1-score	support
geo	0.71	0.60	0.65	705
org	0.40	0.37	0.38	553
gpe	0.54	0.80	0.64	172
tim	0.83	0.55	0.67	244
per	0.66	0.66	0.66	764
art	0.00	0.00	0.00	22
nat	0.00	0.00	0.00	6
eve	0.00	0.00	0.00	8

Description of starter code:

We padded the sentences to have 70 words. There isn't much to the starter code. We did not use any pre trained word embeddings rather we let the word embeddings be learned.

For more work we will use a BiLSTM as well as try to train it with embeddings to see if we fare better.