TEAM 12
CSCI 5832: Natural Language Processing
Eravatee Raje, Ishika Patel, Sukeerth Kalluraya
29 November 2021

Assignment 3: Named Entity Recognition

PROJECT WRITE UP

The project asked students to develop a method to process named entity recognition on a language data set from the biomedical field. The data we are given to train our approach to named entity recognition has been streamlined, and there are only three tag types that we work with for this project. We are looking to tag a final set of data with IOB tags, also known as inside, outside, beginning tags.

Our team worked together to program and debug a Hidden Markov Model Solution to the Named Entity Recognition prompt. Our first step in this process however was not the Markov Model itself. The problem statement itself gave us three unique approaches to sift through. We did preliminary research based on our curiousities of the best solution; each member researched one method thoroughly. We debated between using a feature-based approach similar to assignment 2 and the Hidden Markov Based approach, and landed on the Hidden Markov Model based on the ability to program something new!

Our next relevant step was to understand the Hidden Markov Model as well as features to decode the model's output, like the Viterbi Algorithm. Understanding inputs, the process and the outputs to the Hidden Markov Model were crucial to the actual implementation portion. A really helpful and short paper we looked into to understand the probabilities needed for the model was from the IJNLC -- a citation has been included. Other resources we came across were algorithms to build a Hidden Markov Model as well as even learning about a Python library that can compute the model for you! Although we didn't use this library (haha), it was a really awesome experience to learn about a new python library and its features then take it one step further and implement the model on our own.

During the implementation process, we used pair(s) programming as a method to collaborate and bounce ideas off of one another on what the code would actually look like. We started off with simply sorting the given training data into tuples and splitting the data into training and testing 80/20 sets using the sklearn library. Over Zoom we then commented out some pseudocode + labled sections we needed, and brainstormed the order we needed to implement each piece in. The major parts of the Hidden Markov Model Algorithm include: start probabilities that look at a sentence's tag and the total number of sentences in the corpus, creating bigrams, computing bigram frequency counts in order to fulfil each of the conditional probabilities/Transition Probability (ie total Tags from O to B), Emission Probability of assigning a particular tag to a word, implement some sort of smoothing, and implementing the Viterbi Algorithm to decode what all the number mess means. Another consideration we needed to make in this process is how to handle unknown words. We decided that the best method to handle unknowns is to filter our words with 0 and 1 frequencies.

We pass the various probabilities generated into the Viterbi algorithm to decode the numbers and tag the words. One big consideration in using this model is understanding how to make it a little more efficient. Looping through tuples already gives us an n-squared efficiency. The Viterbi Algorithm is also a very intensive and expensive algorithm to implement. However, this is one of the trade-offs we had to take in order to implement a more dynamic modeling approach to named entity recognition.

All in all, this was a very challenging and interesting implementation of a named entity recognition system. We had to jump through online communication loops and work together to brainstorm and debug virtually via Zoom. This was a full group effort where each of us worked on every aspect: from research, to implementation, to debugging and the report. We are proud of the Hidden Markov Model put forward and eager to learn more about the methods that csn be used to fine-tune it after the fact.

CITATION

Morwal, Sudha and Jahan, Nusrat and Chopra, Deepti, Named Entity Recognition using Hidden Markov Model (HMM) (2012). International Journal on Natural Language Computing (IJNLC) Vol. 1, No.4, December 2012, Available at SSRN: https://ssrn.com/abstract=3758852

NLP Smoothing Tutorial. Stanford University. https://nlp.stanford.edu/~wcmac/papers/20050421-smoothing-tutorial.pdf