CSCI 699: Assignment #1
Named Entity Recognition

Sabyasachee Baruah
5102289620

February 17, 2019

1 CRF

I have used python-crfsuite to train the conditional random field tagger. The following features have been used -

1. word (case insensitive)
2. whether the word is in uppercase, whether it is a digit, whether it is a title
3. word suffix - last 3 characters, last 2 characters
4. part of speech tag
5. part of speech tag prefix - first 2 characters
6. item 1, 2, 4 and 5 for the preceding and succeding word

I followed the hidden markov models chapter [1] of the Speech and Natural Language Processing book, authored by Jurafsky and Martin, to understand first order markov models and the viterbi algorithm. Then I followed [2] to get acquainted with conditional randon fields and how it solves the label bias problem faced by other sequence models. The features selected were inspired by similar work done using conditional random fields on CoNLL 2003 named entity recognition datasets.

2 LSTM

I have used flair package for training a bidirectional lstm model with a conditional random field as the last layer. The flair package has been built on top of pytorch, and provides a very convenient and easy way to train any sequence or classifier model for natural language processing tasks. The following features have been used -

1. word embeddings (glove)
2. character embeddings

I initially tried to train a sequence tagger using PyTorch and word embeddings, however due to time constraints, development set performance was not satisfactory. This was also my first time using PyTorch and I did get the opportunity to learn this wonderful deep learning library. I turned to Flair to accomplish my task more easily. I followed the Loading Corpus and Training tutorial.

The tagger takes an hour to complete each epoch of training, so I could only test with atmost two different sets of features. The first set excluded character embeddings, and the second set includes it. The development set micro F1 score after the first epoch increased by 0.2 when I used character embeddings.

References

[1] Jurafsky, M. Hidden markov models. https://web.stanford.edu/~jurafsky/slp3/A.pdf.

[2] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (2001).