Last NLP Project Report

The project I choose is using featured-based neural network to identify and classification trigger of events in a Chinese database. The project is difficult, which includes lots of aspects such as machine learning, as well as the training data I extracted is very skewed, but I think the project is precise enough for practical use.

The running environment is Anaconda3, pynlpir(Python API for NLPIR), Theano and Keras.

The pipeline of the project can be separated by following parts: First, separating the sentences into words. I used NLPIR to do this. It says this separator performs best among many separators. The performance of the library is pretty good, and won't lead to any error to the final result of pipeline.

Second, POS is from the first NLP project written by myself. This POS program performs good and is a simple process in the project.

Third, the syntax parsing process is also from previous project. The trained model can be directly used in this project, because these two training datasets are similar, both containing short news articles.

Forth, extracting the features. I used featured-based method to predict trigger, so I extract 16 dimensions' vector to predict whether this word is trigger, then use a dictionary to classify the trigger. The feature vector extract following features: the word, POS of the word, the depth in decency tree, left three POS and left one word, right three POS and right one word, dependency type, head word and head POS. Besides, I add "the number of sentences in this text" and "the index of the sentences" these two feature, in order to show a sentence-size relation. For example, it's very low possibility that triggers in the first or the last sentence. However, this two features performs bad, so I put weight of 0.1 to them in the training part.

Fifth, the machine learning part. This is the most difficult part. As I take every word as a sample, the training set becomes very skewed, which contains only 1.5% positive samples. So all I do is to try. After implement neural networks and trying different structure, I find that fully connected layers' model is complex enough for this project.