Eric Minor and Amogh Jahagirdar

In order to correctly tag gene names in sentences, we have made use of the python libraries nltk, sklearn, and seqlearn.

SequenceLearn (seqlearn) is an extension of sklearn that extracts feature information about each word in a sequence passed into it, and uses those features to train a model capable of predicting which tag should be applied to which word. The first task in such an approach is to decide which features to extract from each word in the sequence to maximize the amount of useful data the model has to base predictions off of. The first feature extracted is the word itself. If a word has already been identified as being a gene name in the training data, then it is expected that such a word will be predicted to be gene name. To make more abstract predictions, the two previous words and the next two words are also returned as features, allowing the model to make predictions in reference to the context in which a word appears. All word features are converted to lowercase. In addition, the parts of speech the words in the window are returned using nltk in order to give the model grammatical context. Finally, the current word's casing (upper, lower, or mixed) is returned as a feature. The features extracted from the training data is used to build a StructuredPerceptron model, which is then used for sequence tagging. This approach does lead to a slightly random final model, since the initial model weights before training is random.

Using this approach, F1-measures of 0.60-0.62 are typically obtained. Precision is generally between 0.67 and 0.7 and recall is between 0.55 and 0.58

There are 5 parameters to modify at the start of the identifier.py script.

testFile sets which file you want to predict tags for

testFileAnswers sets which file will be used to evaluate the correctness of the tags. If such a file does not exist, set answers to 0 to skip accuracy measurement and create output.txt only.

trainFile specifies which file will be used to train the model

notConll should be set to 1 if testFile does not contain tags

answers should be set to 0 if you do not want the accuracy of the model to be evaluated.

The file tagged with the model predictions is named output.txt