Name: Zeyuan Li

Andrew ID: zeyuanl

Email: [zeyuanl@cs.cmu.edu](mailto:zeyuanl@cs.cmu.edu)

System Report

1. Please identify/describe any machine learning techniques used.

The system uses conditional random field (CRF) to tag the gene mention in the sentence, since CRF model achieves relatively high performance in the previous years of Gene Mention Tagging task in BioCreative. CRF is a probabilistic model that gives the conditional probability of a possible sequence of T, given the input token sequence O.

The system’s overall precision is 84.6%, recall 62.2% and F-measure 71.7%. The performance of the system is fairly good, taking into account the system uses simple feature set compared to the complicated features sets and pre- and post-processing procedures used by the top-ranked system in the task.

The feature set of the model includes word, orthographic characteristics, but also gene specific feature, like the informative gene-specific substring “gene”, “homeo”. The system also includes prefix and suffix of the word. This is due to the fact that some prefix and suffix are indicator of special type of name entity. There is bracket and punctuation information contained to help us find gene mention. Besides, the system used a window of [-2, 2] of words near to the current word to derive some co-occurrence information about some pair of gene mention sequence.

1. Please identify/describe any NLP techniques/components used.

Gene Mention Tagger system uses Stanford NLP tokenizer for sentence tokenization and Mallet’s sequence tagging CRF machine learning model.

Stanford tokenizer divides text into a sequence of tokens, which roughly correspond to "words". In the system, Stanford tokenizer helps to tokenize the input sentence into tokens, which is crucial to the tagging task in next step.

Mallet is a MAchine Learning for LanguagE Toolkit for NLP, document classification, sequence tagging etc. developed by UMass. In specific, the system uses only the sequence tagging part of the package for labeling the test sentence whether each tokens belongs to B (begin of gene mention), I (intermediate gene mention) or O (not belongs to a gene mention).

1. Please identify/describe any external (marked up text) training data used.

The training data used is the first 10,000 of the original “sample.in” input file. These sentences include 67% of the entire dataset, which is a fair partition between training and testing data.

1. Please identify/describe any external lexical resources (terminology lists)used.

Some external lexical resources are not used in the system, like gene related lexical database and web services. Incorporating this kind of material, I believe, could further improve the system performance.

1. Please describe any rule sets used.

The system do not use any rule set base method, rather it uses the probabilistic view of the gene tagging sequence, regarding the whole process as a maximum likelihood estimate of the tagging sequence given the observed tagging sequence in training data.

1. System performance

The system’s overall precision is 84.6%, recall 62.2% and F-measure 71.7%. The performance of the system is fairly good, taking into account the system uses simple feature set compared to the complicated features sets and pre- and post-processing procedures used by the top-ranked system in the task.

1. Please describe the general data flow in your system.

Gene Mention CPE engine is consists of three parts, a Collection Reader, some Analysis Engines and a Collection Reader.

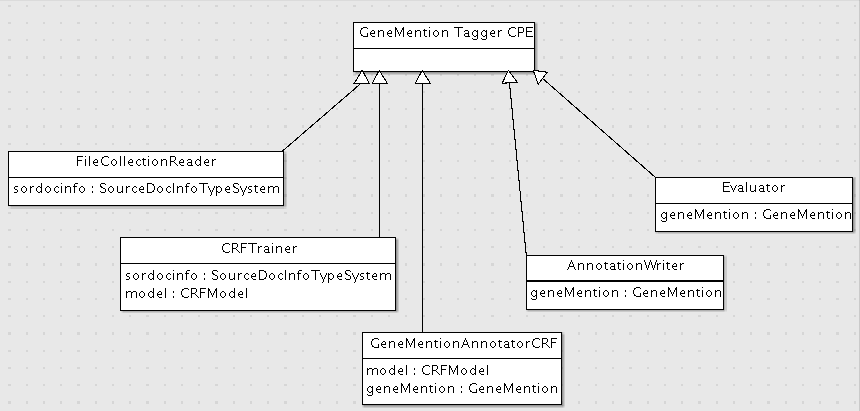
The Collection Reader is named “FileSystemCollectionReader”, which reads input document into Type System “SourceDocInfoTypeSystem” and passes it to CAS.

The Analysis Engines consists of two components, namely “TrainCRF” and “GeneMentionAnnotatorCRF”, which the former trains a CRF model and the latter do the sequence tagging of the test data. Type System “CRFModel” passes the trained model to annotator on only path attribute. The annotated gene mention is stored in Type System “GeneMetion”.

Finally, the CAS Consumer includes two parts: “Evaluator” and “AnnotationWriter”, which the former is to calculate precision, recall and f-measure and the latter is simply output the gene mention result to the file.

1. System UML

The system UML domain diagram is showed blow. The type system is in the attribute of each domain object.



1. Design Pattern

The system uses Information Expert design pattern. For example, the FileCollectionReader should know the input file. Therefore assign read() process in FileCollectionReader class. The training process of the CRF model should be the information expert of the training process and assign the responsibility to the CRFTrainer process.