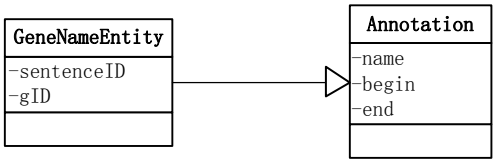
Short Summary for hw1

Ke Xu (kex)

1. Type System



sentenceID stores the prefix-string that identifies a sentence.

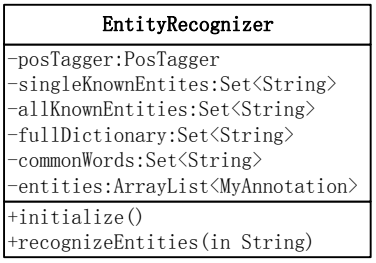
gID stores the global number for each sentence.

2. Collection Reader

It accepts 3 configuration parameters : inputDirectory, Encoding, Language, and converts each input file to a string, then add the string into the CAS instance.

3. Annotator

It processes one input string per time, and add analyzed results into input CAS instance. The main processing routine is encapsulated in *EntityRecognizer* class.



The main routine for recognizing is as follows:

1) Building Dictionaries
2) Preprocessing
3) Searching
4) Post processing

Building Dictionaries :

There are 3 kinds of dictionaries used.

First, the *Gene Name Dictionary*. I built this dictionary by combining several existed gene dictionaries. Items in the dictionary are not limited to single word, but multi-word terms. It contains about 700,000 terms in total.

Second, the *Common Word Dictionary*. I built this dictionary by collecting words from *Gone With the Wind, Shakespeare Plays* and other literatures, where gene name is unlikely to appear. I consider this dictionary as a large stop-word list, which was used at the preprocessing step.

Third, the *Full English Dictionary*. This is a 170,000 words English dictionary used for reference at the post processing step.

Preprocessing :

For each sentence to be handled, first I converted them to tokens with Stanford core NLP, then I delete 'common words' with the *Common Word Dictionary*.

Searching :

For the remaining words, I simply looked up each words in the *Gene Name Dictionary*; and I combine continuous candidates to one candidate.

Post processing :

For those single-word candidates, I looked up them and the stemmed terms (generated by Stanford core NLP, Morphology class) in the *Full English Dictionary*, and simply dropped those appearing in the dictionary.

And for each candidate, I used Stanford core NLP to tag the Part-Of-Speech for each words. Only nouns were maintained.

4. Consumer

It simple generates the required format of gene names.

5. Performance

Take the *sample.out* as standard results, only considering cases that are perfectly match *sample.out* (including length and order) , then the *precision* is about 0.35, *recall* is about 0.32.