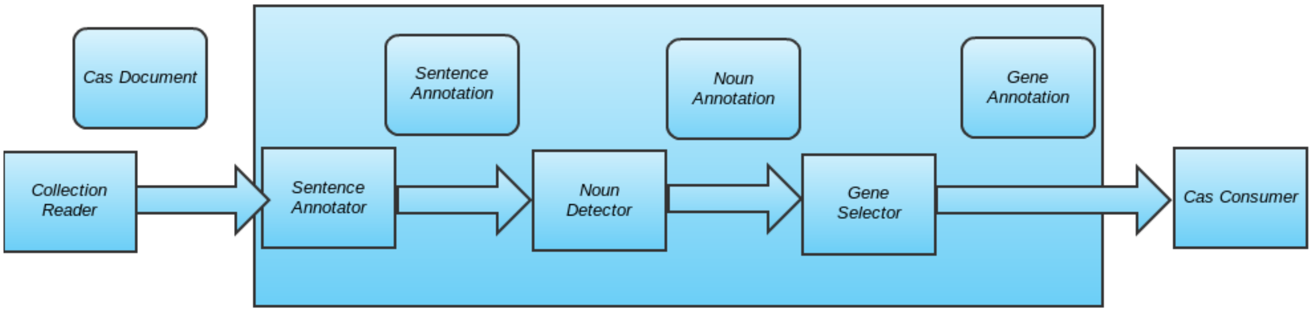
Gene Annotation using UIMA framework report

Li Yiwei

The use of named entity recognition is the essential step for many language processing functionalities like Question Answering. This project uses UIMA framework to implement gene annotation functionality.

Definitely there are many ways to implement the gene annotations. The architecture of the pipeline that is used in the UIMA system is quite clear in my design which is shown as below:



Gene Annotation Analysis Engine

The Analysis Engine is composed of three annotators, namely the Sentence Annotator, Noun Detector and Gene Selector.

The Sentence Annotator takes the input from the collection reader which is the part of cpe system. It will truncate the original document into independent sentences which are separated by line separator. Each sentence annotation has a sentenceID parameter to keep track of its identity and location.

The second part of the AE in my design is a Noun Detector. It is mainly dealing with the job of tokenization and POS tagging. The tokens tagged as nouns are selected as the Noun Annotations. The consecutive nouns will be combined as the Noun Phrases which is one Noun Annotation.

The last part is the Gene Selector. It will take the Noun Annotations as input and will use some techniques to determine whether or not the Noun Annotation is a gene name. If so, it will be recognized as a Gene Annotation. This is the most difficult and flexible part of the project in my design as there are many approaches of doing it. Due to the limited knowledge of machine learning and language processing technologies, I tend to use the most basic method of dictionary look-up.

In order to enable the system to recognize from a list of nouns the gene names, I need to find a database of the gene names to do the dictionary look-up. The database can be online from a website or can be locally saved in a file or database system. After trying both approaches, the main problem that I found by using some of the online databases is the access speed. The processing time is strictly limited by the access time to the internet and can very geometrically and timely. It needs for about half a hour in order to do the named entity recognition of gene concept for 1000 sentences which is quite a big drawback of the system. Though some techniques are used to improve the system like the historical data will be used for the second-time search, the speed is still not very promising. I also found some website they provide the API for accessing their remote database by using SOAP and WSDL which is faster than simply doing HTML analysis of the webpages that fetched from a Get or POST message to the server. This will be likely the direction that I will work on by using SOAP and WSDL in the future's implementation and research work.

I finally chose to use a locally stored database to do the dictionary look-up which has the advantage of fast speed. But its accuracy and flexibility is not very promising.

I will also try to do more research on statistical method of dealing with named entity recognition as the work for next step. I noticed that there are some API available on the internet like lingpipe which provide implementation for many of the basic NLP techniques. So the implementation of Gene Selector can also be done by using the lingpipe API like its Nbest and confidencial N Best HMM models. And there are pre-trained models available directly for gene tagging in the lingpipe website which is very friendly to use.