hw2 report for 11791

andrewID: xiaoxul

October 10, 2014

1 System design

- Reader does the job of reading the text from input file.
- Annotator does the job of processing the text and recognize the entities, which are the names of genes in this homework.
- Consumer does the job of writing the result sets into output file. (I also include codes comparing the sample output and my output to calculate the precision, recall and F-measure before the grader came out. To make sure that the grader does its job without any accidents, I commented them all.)

1.1 Type system

1. **Sentence**
   - ID: imply the ID of the sentence
   - Content: imply the ID of the sentence
2. **Genetag**
   - ID: imply the ID of the sentence where the Genetag is recognized.
   - Content: imply the name of the Genetag.
   - Begin: imply the start index of the Genetag in the sentence.
   - End: imply the end index of the Genetag in the sentence.
   - confidence: imply confidence of the entity picked out by lingpipe.
   - casProcessorID: imply which NER generate this gene tag.
3. **Gene**
   - ID: the same as Genetag.
   - Content: the same as Genetag.
   - Begin: the same as Genetag.
   - End: the same as Genetag.

1.2 Annotator

1. sentence annotator
   - This annotator is implemented by ***Sentence_Annotator.java***.
   - Its process function splits the text get from input file into sentences.
   - Mark the ID and Content of Sentence.
2. gene annotator using lingpipe

   - This annotator is implemented by ***Geneannotator_lingp.java***.
   - Its process function gets the sentences and uses lingpipe's **nBestChunk** to recognize the gene names in each sentence.
   - Mark the ID, Content, Begin and End, as well as Confidence and casProcessorID, which is lingp, of Genetag.

3. gene annotator using abner
   - This annotator is implemented by ***Geneannotator_abner.java***.
   - Its process function gets the Sentences and uses abner's **tagger** and **getEntities** to recognize the gene names in each sentence.
   - Mark the ID, Content, Begin and End, as well as casProcessorID, which is abner, of Genetag.

4. gene decider
   - This annotator is implemented by ***Annotator_Decider.java***.
   - Its function process gets the Genetags and does a decision-making job. All the Genetags from lingpipe, those whose confidence is greater than 0.2 will be directly given to new type system, while those confidence is less than 0.2 will be evaluated whether it is in the Genetags from abner. On the other side, all the Genetags from abner will be given to the new type system. (Hate to admit that this is not intelligent at all , but it gets beautiful performance score with the sample output.)
   - Mark the ID, Content, Begin and End of Gene, which is the type system for consumer to output .

2 Performance evaluation

Precision: 0.7294874173
Recall: 0.64593484807
F1 Score: 0.685173355015
By grader from *https : //github.com/amaiberg/software − engineering − preliminary/tree/master/ grading_hw1_2*.

3 External information

The entity recognizer I used:
1. *lingpipe*: *http : //alias − i.com/lingpipe/index.html*
2. *abner*: *http : //pages.cs.wisc.edu/ bsettles/abner/*
The API:
3. *UIMA API (2.6.0)*: *https : //uima.apache.org/d/uimaj − 2.6.0/apidocs/index.html*
4. *JAVA API (SE8)*: *http : //docs.oracle.com/javase/8/docs/api/*