## Report of Named Entity Recognizer

**Haohan Wang**

**ID: haohanw**

This is the report of the design of the homework Named Entity Recognizer with UIMA SDK, this report is mainly composed of three parts: the architecture design, the algorithm design and the evaluation of result.

### Architecture Design

#### Main Architecture & Pipeline:

The pipeline mainly contains three parts, the reader of the file, the Analysis Engine to extract the genetags and the consumer to output the result.

The AE uses a pre-trained database to extract the information of input sentences.

The system also keeps the results with different factors and evaluates the results to select the best threshold for future use.

#### Type System



The type system contains only one Type System, the genementiontype, it contains four variables. GeneDoc is a String contains the line of input except the line ID. LineID is a string containing the line ID. StartOffset and EndOffset are Integers for the start offset and end offset of the target words before spaces are eliminated.

### Algorithm Design

#### FileReader

The reading system is mainly based on the UIMA example.

Read one line of the input and generate the CAS for AE.

#### Analysis Engine

**LingPipe HMM**

Analysis Engine is based on the n-best words with confidence of LingPipe (Add Maven Dependency of alias.com lingpipe). The named entity recognition of it mainly involves the supervised training of Hidden Markov Models.

**External Model**

A pre-trained HMM for genetags named ne-en-bio-genetag.HmmChunker (<http://alias-i.com/lingpipe/web/models.html> ) is imported and used for the algorithm. The Pre-trained model is imported with this function:

*ConfidenceChunker chunker*

*= (ConfidenceChunker) AbstractExternalizable.readObject(modelFile);*

**Name Entity Recognizer**

The main part of this AE is a function:

*Iterator<Chunk> it*

*= chunker.nBestChunks(cs,0,cs.length,MAX\_N\_BEST\_CHUNKS);*

This function receive the input char[], with the parameters of start, end, and the number of the best matching results.

This function returns a Chunk for the list, containing the information of the matched word: start, end, and confidence, etc.

**Filter the result**

Only the name entity with a confidence greater than a threshold is added into the annotator.

For each Name Entity, the annotator contains the information of its lineID, its start-offset, its end-offset and the whole line text.

#### Consumer:

The Consumer is mainly based on the UIMA example.

Before output the file, the count of spaces should be eliminated for the start and end of each word.

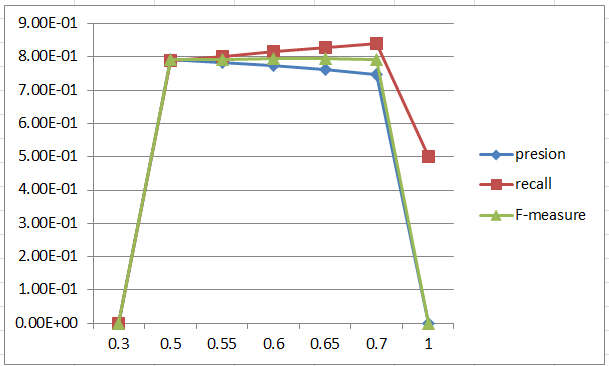
### Evaluation of Result

After several round of experiments, the result is the best when the threshold of confidence is 0.6. The precision is 0.773805, recall is 0.815323, F-measure is 0.794022

The following chart shows some interesting numbers from the experiments:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.3 | 0.5 | 0.55 | 0.6 | 0.65 | 0.7 | 1 |
| presion | 5.75E-05 | 0.79275 | 0.783661 | 0.773805 | 0.761266 | 0.74692 | 0.000109511 |
| recall | 4.83E-05 | 0.788519 | 0.801748 | 0.815323 | 0.828546 | 0.840843 | 0.5 |
| F-measure | 5.13E-05 | 0.790629 | 0.792601 | 0.794022 | 0.793482 | 0.791104 | 0.000218974 |

The following diagram shows that the threshold should be around 0.5-0.7:



The following diagram shows why we choose 0.6 as threshold:

