****

**SOEN 691B Semantic Computing**

**Winter 2013/2014**

**Assignment #1**

**Report**

**Presented by: Anish Chavan**

**Studentid :6421180**

**Problem Defination:**

To develop an Information Extraction (IE) system for entities and relations in the university domain.

**Pipeline :name(anish)**

**using japerules**

The pipeline used in the assignment consists of inbuilt gate resources,plus additional developed resources using Java Annotation Patterns Engine .This engine provides predefined rules which can be used to create new system and add to the exixting system.

The following is the graphical and detail description of the pipeline used in the system.

Jape Transducer

output-(Has\_position)

Jape Transducer

output-(universityposition)

Jape Transducer

output-(OrginizationUnit)

Jape Transducer

output-(Belongs\_to)

Jape Transducer

output-(University Person)

Jape Transducer

output-(OrginizationUnit)

Jape Transducer-Output- (university)

ANNIEPlugin

ANNIE OrthoMatcher

Custom

gazetteer

ANNIEPlugin

ANNIE Sentence Splitter

ANNIEPlugin

ANNIE English Tokeniser

ANNIEPlugin

Document Reset PR

ANNIEPlugin

ANNIEPlugin

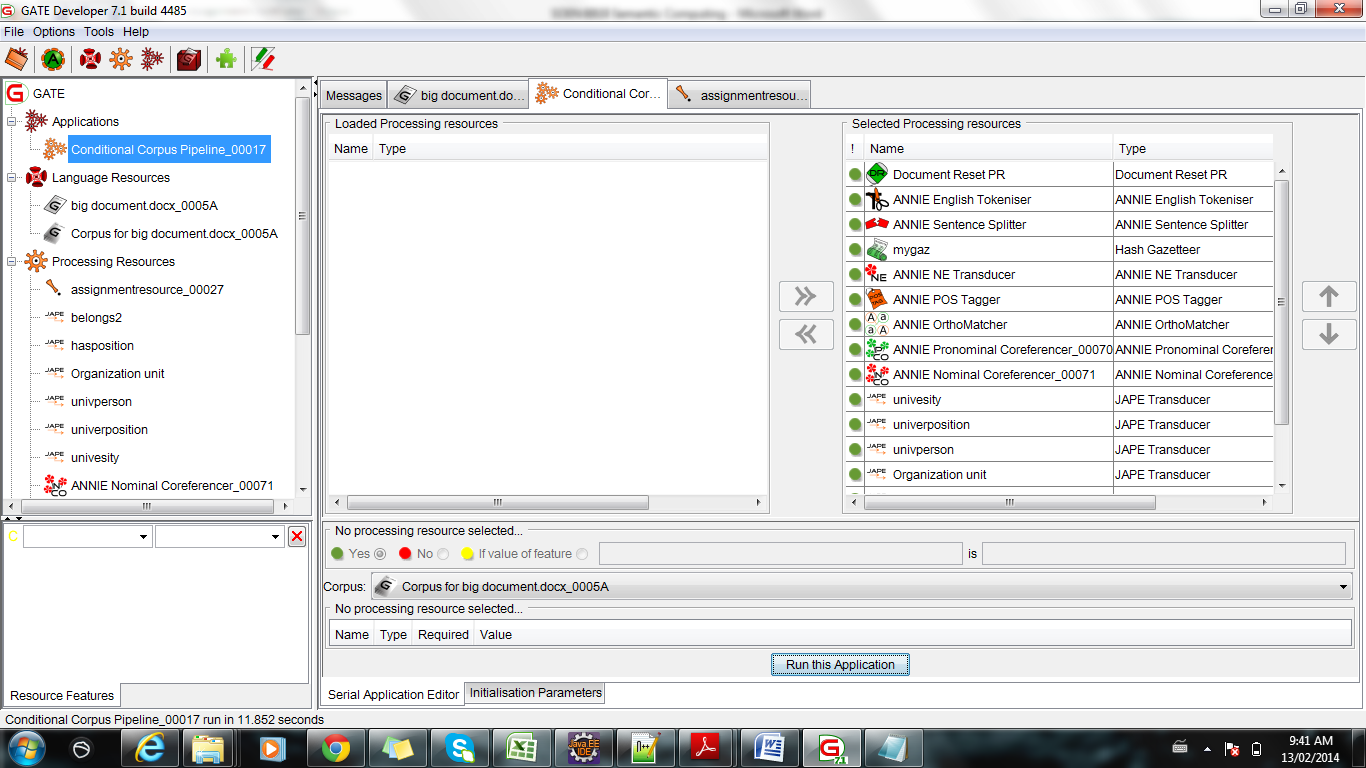
ANNIE NE Transducer

ANNIEPlugin

ANNIE POS Tagger

CustomPR to write the output to the html

**ScreenShot:**



**Gazetteer:**

The following are the list of gazzetter which actually used for university domain.

**Source**: The following gazetteer list was developed from annie.Annies gazetteer was modified to match the university domain.

**Structure**(actual gazetter was big but only following ver relevant to domain of university)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| city.lst | location | city |  | Lookup |
| city\_cap.lst | location | city |  | Lookup |
| country.lst | location | country |  | Lookup |
| country\_abbrev.lst | location | country\_abbrev |  | Lookup |
| country\_adj.lst | country\_adj |  |  | Lookup |
| country\_cap.lst | location | country |  | Lookup |
| currency\_prefix.lst | currency\_unit | pre\_amount |  | Lookup |
| currency\_unit.lst | currency\_unit | post\_amount |  | Lookup |
| department.lst | organization | department |  | Lookup |
| dept.lst | organization | department |  | Lookup |
| facility.lst | facility | building |  | Lookup |
| facility\_key.lst | facility\_key |  |  | Lookup |
| facility\_key\_ext.lst | facility\_key\_ext |  |  | Lookup |
| jobtitles.lst | jobtitle | jobtitles |  | Lookup |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| person\_ambig.lst | person\_first | ambig |  | Lookup |
| person\_ending.lst | person\_ending |  |  | Lookup |
| person\_female.lst | person\_first | female |  | Lookup |
| person\_female\_cap.lst | person\_first | female |  | Lookup |
| person\_full.lst | person\_full |  |  | Lookup |
| person\_male.lst | person\_first | male |  | Lookup |
| person\_male\_cap.lst | person\_first | male |  | Lookup |
| person\_relig.lst | person\_full | relig |  | Lookup |
| person\_sci.lst | person\_full | sci |  | Lookup |
| person\_spur.lst | spur |  |  | Lookup |
| province.lst | location | province |  | Lookup |
| racecourse.lst | location | racecourse |  | Lookup |
| region.lst | location | region |  | Lookup |
| region\_cap.lst | location | region |  | Lookup |
| region\_uk.lst | location | region |  | Lookup |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| universities.lst | univ | univ |  | Lookup |

**Size:**

Approximate size of the list is around 1 kb to 4 kb.

Size of university.list=226KB

jobtitles.list=108KB.

The gazetteer consist of def.list file which acts a index for the jape grammar to match the tex

**Named entity recognition :**

**Annie ProcessingResources:**

The pipeline designed consists of Annie transducers like Named Entity transducer. Which gives entities like Person ,location and date etc...

The other processing resources are like Parts of speech tagger ,which is responsible for identifying parts of speech and adds to the root of the word as the feature .

**Custom jape transducer:**

The main entities which were recognized where university names, university position ,universityperson,organization unit.

The jape rules consists of L.H.S and R.H.S.The L.H.S are the condition matched in the text when the mathc is done R.H.S is executed

**university.jape:**

**L.H.S--**

(

{Token.string == "University"}

({SpaceToken.kind!="control"})+

({Token.string == "of"})

({SpaceToken.kind!="control"})+

(

({Lookup.majorType==location})\*:cityname|({Token.category==NNP})\*|{Token})+

):univname2

**R.H.S**

When above rule is matched university annotationis created.

**UniversityPosition.jape**

Rule: UniversityPsoition

(({Token.string=="vice"}|{Token.string=="Vice-"}|{Token.string=="Vice"}|{Lookup.majorType == jobtitle})?

{Lookup.majorType == jobtitle}

(

{Lookup.majorType == jobtitle}

)\*

)

**R.H.S**

When above rule is matched UniversityPosition annotationset is created.

**OrganizationUnit.jape**

**L.H.S**

Options: control = appelt

Input: Token Lookup

Rule: OrganizationalUnit1

Priority: 55

( {Lookup.majorType ==organization}

(

{Lookup.majorType ==organization}

))

:OU

-->

:OU.OrganizationalUnit={kind="OU",rule="OrganizationalUnit"}

**UniversityPerson.jape**

Phase: UniversityPerson

Input: UniversityPosition Person Split

Options: control = appelt

Rule: UniversityPerson1

(((({Person}):P1

({UniversityPosition}) : UP1

({Split})

):condition1|

(({UniversityPosition}):UP2

({Person}):P2

({Split})

):condition2

)|(

({Person}):P3

({Split})\*

({UniversityPosition}):UP3

):condition3

|(

({UniversityPosition}):UP4

({Split})\*({Person}):P4):condition4):Univpos

**Other components:**

The other components used or developedare as follows

**Relation extraction:**

This included identifying the relation between the university and university person.and also person lies at what position.

To identify the relation between the person and the entity the distance was considered to be the main parameter.

for example

(person1) (university) .....person2 then the mininmum of both the persons distance was calculated from the entity.

smaller the distance more is the possibility of the person lying in that entity

Rule: getz In the R.h.s The offset value was retrieved using

( java and minimum distacne ws calculated

Hashtable<Long,Long> h=new Hashtable<Long,Long>();gate.AnnotationSet university1=(gate.AnnotationSet)bindings.get("univ");

Annotation UIterator=university1.iterator().next();

long number=UIterator.getStartNode().getOffset();

gate.AnnotationSet Person1=(gate.AnnotationSet)bindings.get("ztag");

Annotation P1Iterator=Person1.iterator().next();

AnnotationSet temp=Person1.get("UniversityPerson");

for(Annotation annot: temp){long map = annot.getStartNode().getOffset()

long Result=map - number;

if (Result < 0)

{

Result = -Result;

}

long key = annot.getStartNode().getId();

h.put(key, Result);System.out.println(Result);}

System.out.println("============================================================");long min=0;

long x=0;for(Long k:h.keySet()){

long temp2= h.get(k); if (min==0)

{min=temp2;

x=k;

System.out.println("K"+k);

}

else if(temp2<min)

{

min=temp2;

System.out.println("K"+k);

x=k;

//final=k;

}

(({UniversityPerson})?

({Split})\*

({UniversityPosition}):univ

({Split})\*

({UniversityPerson})?

{Split}

)|(({UniversityPerson})?

({Split})\*

({UniversityPosition}):univ

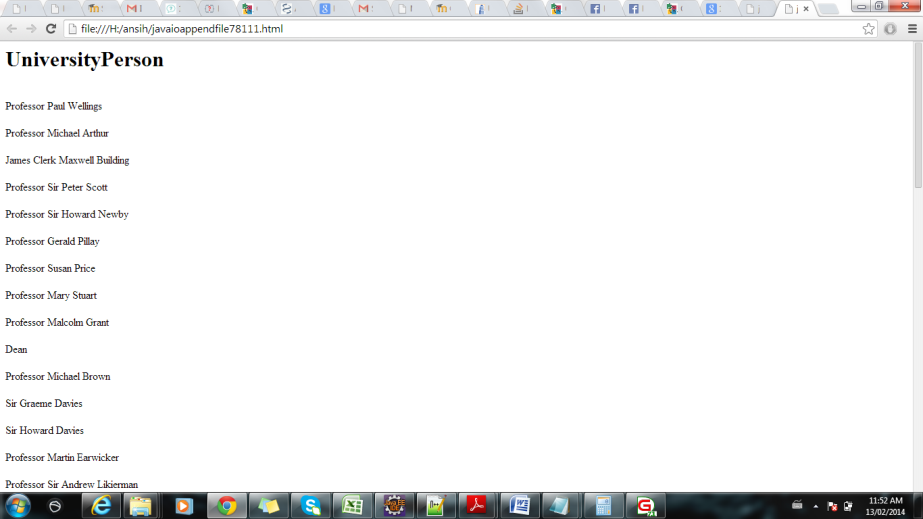
({Split})\*

({UniversityPerson})?

({Split})\*

)

**HTML export** was done by creating new Processing resource in java and was added at the end of the



pipeline. The zip file contains the demo run on the five test documents

**Co-refrence:**

The corefrence relation was done using Annie nominal and pronominal transducer.

The result from the data was used in order to obtain all the matches like[5626,28282..]

and only out put the non redundent data.

The co-reference algorithm was implemented by using the matches value obtained from pronominal Annie tranceducer..The matches were found of the ids in java using following lines of code.

for(Annotation annot: blacklist1)

{ int bb1=0;

FeatureMap map = annot.getFeatures();

if((map.get("matches"))!=null)

{annotCategory=((ArrayList<Long>) map.get("matches"));

for(Object off:annotCategory)

{if(bb1!=0){

if(!list.contains((Integer)off))

{list.add((Integer)off);

}System.out.println(list.size());

bb1++......................................These matches are used to collapse the multiple refrence of an annotation.

**Evaluation:**

size of the corpus ins tokens is about 16511 tokens of 5 documents.

following were the results Annotationdiff tool on the 5 documents selected .

(for complete corpus)

* **university-**annotationusing annodiffui

Recall: 0.6389  
Precision: 0.4694  
F-measure: 0.5412  
  
Correct: 23  
Partially: 10  
Missing: 3  
False positives: 16

**Quality Assurance:(corpus statictis)**

**Annotation match onlyA onlyB overlap precB/A Recall F-measure**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Belongs\_to | 0 | 0 | 1706 | 0 | 0.00 | 1.00 | 0.00 |
| Has\_Position | 0 | 0 | 690 | 0 | 0.00 | 1.00 | 0.00 |
| Organizationalunit | 0 | 19 | 0 | 0 | 1.00 | 0.00 | 0.00 |
| university | 103 | 25 | 358 | 102 | 0.36 | 0.89 | 0.52 |
| UniversityPerson | 0 | 0 | 324 | 0 | 0.00 | 1.00 | 0.00 |
| universityposition | 0 | 311 | 0 | 0 | 1.00 | 0.00 | 0.00 |
| UniversityPosition | 0 | 0 | 541 | 0 | 0.00 | 1.00 | 0.00 |
| Macro summary |  |  |  |  | 0.34 | 0.70 | 0.07 |
| Micro summary | 103 | 355 | 3619 | 102 | 0.05 | 0.37 | 0.09 |
| Belongs\_to | 0 | 0 | 1706 | 0 | 0.00 | 1.00 | 0.00 |
| Has\_Position | 0 | 0 | 690 | 0 | 0.00 | 1.00 | 0.00 |
| Organizationalunit | 0 | 19 | 0 | 0 | 1.00 | 0.00 | 0.00 |
| university | 103 | 25 | 358 | 102 | 0.36 | 0.89 | 0.52 |
| UniversityPerson | 0 | 0 | 324 | 0 | 0.00 | 1.00 | 0.00 |
| universityposition | 0 | 311 | 0 | 0 | 1.00 | 0.00 | 0.00 |
| UniversityPosition | 0 | 0 | 541 | 0 | 0.00 | 1.00 | 0.00 |
| Macro summary |  |  |  |  | 0.34 | 0.70 | 0.07 |
| Micro summary | 103 | 355 | 3619 | 102 | 0.05 | 0.37 | 0.09 |

**Document statictics**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Administrators.docx\_00034 | 10 | 69 | 625 | 17 | 0.04 | 0.28 | 0.07 |
| AtoD.docx\_00035 | 23 | 67 | 401 | 10 | 0.08 | 0.33 | 0.12 |
| EtoM.docx\_00036 | 22 | 107 | 447 | 15 | 0.08 | 0.26 | 0.12 |
| RtoZ.docx\_00037 | 19 | 75 | 477 | 18 | 0.07 | 0.33 | 0.12 |
| big document.docx\_00038 | 29 | 37 | 706 | 42 | 0.09 | 0.66 | 0.16 |
| doc2.docx\_00039 | 0 | 0 | 963 | 0 | 0.00 | 1.00 | 0.00 |
| Macro summary |  |  |  |  | 0.06 | 0.48 | 0.10 |
| Micro summary | 103 | 355 | 3619 | 102 | 0.05 | 0.37 | 0.09 |