Assignment #2: Web Page Clustering of Ambiguous Names

Dr. Zornitsa Kozareva
USC Information Sciences Institute
Spring 2013

Task Description: Finding information about people, organizations and locations in the World-Wide-Web is one of the most common activities of Internet users. Because names are highly ambiguous often the returned results are a mixture of Web pages about different people/locations/organizations that share the same name. This assignment concerns the automated identification of ambiguous person names and the grouping together of the returned Web pages into clusters such that each cluster corresponds to the same individual.

You will be given as input a set of web pages returned by a web search engine when a given person name was issued as a query. For the current assignment, we will be only using the first one hundred documents retrieved by the search engine.

As output your system must produce for each ambiguous name clusters of the 100 web pages, such that each cluster corresponds to only one individual.

The gold standard (truth) clusters of the Web pages will be provided from the very beginning.

Your goals are: (1) to find the best clustering solution and (2) to compare and contrast at least one of the following scenarios:

- Scenario I: How different features (unigrams, bigrams, co-occurrence, tf*idf, among others) affect the same clustering algorithm
- Scenario II: How the same feature (unigrams, bigrams, co-occurrence, tf*idf, among others) affects the two different clustering algorithms
- Scenario III: How two different feature sets (unigrams, bigrams, co-occurrence, tf*idf, among others) affect two different clustering algorithms

For this assignment you can use any clustering algorithm (k-means, bottom-up, top-down) either from Weka or from other toolkits you are familiar with. You can use existing Latent Semantic Analysis packages like Infomap, or Latent Dirichlet Allocation packages like Mallet to cluster the web pages.

What to turn in:
Send an e-mail to

Zornitsa Kozareva: kozareva@isi.edu
and
Victor Barres: victor.barres@gmail.com

with subject line **CS544 Homework 2**

In the email attach a zip or tar.gz file named FirstNameStudent_LastNameStudent.zip or FirstNameStudent_LastNameStudent.tar.gz which when unzipped should create a folder **FirstNameStudent_LastNameStudent**

Inside the folder **FirstNameStudent_LastNameStudent** there must be four subfolders

\-- **FirstNameStudent_LastNameStudent_Report**
//should contain only a pdf of the report
// pdf should be named FirstNameStudent_LastNameStudent_Report.pdf
\-- **FirstNameStudent_LastNameStudent_System**
//should contain all your code, tools you have used, scripts
// README file describing how to run the system
\-- **FirstNameStudent_LastNameStudent_CL**
//should contain the produced clusters for each ambiguous name
\-- **FirstNameStudent_LastNameStudent_Score**
//should contain the current scores of your system produced by the evaluation script

Instructions on how to generate the output clustering format for each person name

- Abby_Watkins
- Cathie_Ely
- Dan_Rhone
- Jane_Hunter
- Michael_Howard
- Thomas_Baker
- Tim_Whisler

Where the format of the clustering output should be the same as the Gold Standard format described in the Data Description section. The data output for each person name set should be created in one separate file. The file name should be the person name (blanks replaced by "_") with the ".clust.xml" extension.
Don't forget that all "*.clust.xml" files should be put in the folder **FirstNameStudent_LastNameStudent_CL**

In the folder **FirstNameStudent_LastNameStudent_System** provide us with your source code and a README explanation on how we can run your code to check if it is working (input file/output file). What each program does and what is the sequence in which the things you have programmed should be executed. For readability put comments in your code.

The input file you used for the clustering and what was the program you used for clustering, such that if we have to run it and test what the final clusters are we can do it

The folder **FirstNameStudent_LastNameStudent_Report** should have a pdf version of your report on which you must put the name of your system and your name. This report should be a brief description of your system explaining:

- Used features
- Used clustering algorithms
- Used toolkits and links to the places from wherever you have downloaded them
- Results on FMeasure_0.5_BEP-BER from the evaluation script we provide you for each name and overall
- A comparative study on one of the Scenarios I, II, III or IV and the obtained FMeasure_0.5_BEP-BER corresponding to the Scenario

- An error analysis on the wrongly clustered Web pages or an explanation on why some settings worked better than others
- Optional to submit a suggestion on how to improve the task, or a suggestion on how would you have defined the task if you were the first person to come up with it

In the folder **FirstNameStudent_LastNameStudent_Score** leave the current scores produced by the evaluation scorer. Name your files as:

FirstNameStudent_LastNameStudent_System1
FirstNameStudent_LastNameStudent_System2
FirstNameStudent_LastNameStudent_BestSystem

Where **FirstNameStudent_LastNameStudent_System1** are the scores produced by your first system, which can be for example LSA

FirstNameStudent_LastNameStudent_System2 are the scores produced by your second system, which can be for example k-Means weka unigrams

FirstNameStudent_LastNameStudent_BestSystem are the scores produced by the best setting you have found

Note all these scores should be when all ambiguous names were evaluated.

Timeline:

Date Data Set and Assignment Given:	February 26, 2013
Date Assignment Due:	March 15, 2013
Technical Report Due:	March 15, 2013

Evaluation is based on the:

- FMeasure_0.5_BEP-PER score produced by the evaluation script for all disambiguated and clustered names
- Designed and used features
- **Comparative study between the clustering algorithms**
- Quality of the technical report
- Error analysis and suggestions for improving the task
- How well you beat the ALL-IN-ONE and ONE-IN-ONE clustering baselines

Data Description:
You are provided with two folders that have the following structure

webps
\-- **web_pages** //raw web pages downloaded for each name
\-- **truth_files** //human clustering of the documents for each name

scorer_1.1
//documentation, source and jar files of the evaluation package

In **\--web_pages**, you fill find for each name up to 100 web pages, which were returned from Yahoo! when the person name was issued as a Web query. The pages contain the original formatting (html, xml), which must be cleaned prior to the clustering processing.

In **\--truth_files**, you fill find the Gold Standard (i.e. the true clustering) for each person name. The Gold Standard files are named "person_name.clust.xml".
Each file contains a root element "<clustering>" followed by one "entity" element for each entity. The entity element has an identifier attribute ("id") with an integer value. Nested in the "entity" element there are "doc" elements (pages that refer to this particular entity), each of which has a "rank" attribute that corresponds to the ranking information provided in the xml file described above.

Note that a document might have been clustered in more than one entity. This is the case when multiple person names referring to different entities appear in a single document. Also, note that a person name may have a namesake that is not a person (for instance an organization or a location). In those cases the non-person entity will have its own cluster. Finally, when the annotator could not cluster a page it was included under a "discarded" element. The reasons for this might be the non-occurrence of the person name in the page (probably because Yahoo index had outdated information when the corpus was built) or simply that the human annotator could not decide whether to cluster that page. Discarded pages are not taken into account for the evaluation.

Here is an example of what the gold standard files looks like:

```
<clustering>
   <entity id="0">
      <doc rank="0"/>
      <doc rank="5"/>
   </entity>
   <entity id="1">
      <doc rank="1"/>
      <doc rank="3"/>
          <doc rank="5"/>
      <doc rank="10"/>
   </entity>
   ...
   <discarded>
      <doc rank="8"/>
      <doc rank="9"/>
   </discarded>
 </clustering>
```

Note that empty lines are permitted. Space and tab do not have special meaning in the file.

Some files that appear in the list of downloaded documents might contain no text, probably because it was not possible to download them. Those pages where not clustered by the human annotators and should not appear in the "clust.xml" files.

Scoring Package

The scoring folder includes the jar file with the source code and a basic documentation.

Any suggestions, improvements or new measures are very welcomed (write to Javier Artiles javart@gmail.com who is the organizer of the Web People Search Task for 2007, 2009 and 2010).

This program scores the performance of one or more systems according to several optional evaluation measures.

USAGE:
java SystemScorer [keysDir] [systemsDir] [outputDir] [MEASURES] [BASELINES] [OPTIONS]

[keysDir] Directory containing all the gold standard for the clustering problems.
Files must be well formed XML, follow the WePS 2007 clustering format
and filenames end in 'clust.xml'.

[systemsDir] Directory containing all the systems solutions to evaluate using the
following structure

systemsDir/TEAM_A/problem1.clust.xml
systemsDir/TEAM_A/problem2.clust.xml
systemsDir/TEAM_A/...
systemsDir/TEAM_B/problem1.clust.xml
systemsDir/TEAM_B/problem2.clust.xml
systemsDir/TEAM_B/...
systemsDir/...

[outputDir] Directory where all the results will be written

MEASURES:
-ALLMEASURES Evaluates all the available measures
-P Purity
-IP Inverse purity
-FMeasure_0.5_P-IP F-measure for Purity and Inverse Purity (alpha=0.5)
-BER BCubed Recall (extended for multiclass problems)
-BEP BCubed Precision (extended for multiclass problems)
-FMeasure_0.5_BER-BEP F-measure for BCubed Precision and Recall (alpha=0.5)
-PR Pairs measure using Rand Statistic
-PJ Pairs measure using Jaccard Coefficient
-PF Pairs measure using Folkes and Mallows

BASELINES:
-AllInOne
-OneInOne
-Combined

OPTIONS:
-overwrite overwrites previous evaluation files (.eval) if necessary.
-average prints the averaged scores for all the teams

EXAMPLE (using the official annotation set as key and also as a team, evaluating baseline answers):

$ /usr/lib/jvm/java-6-sun-1.6.0.03/bin/java -cp distributions/1.1/wepsEvaluation.jar es.nlp.uned.weps.evaluation.SystemScorer weps07test/truth_official/ weps07test/test_system/ tmp -ALLMEASURES -AllInOne -OneInOne -Combined -average

WePS 2007 Evaluation Package (http://nlp.uned.es/weps)

Key clustering files path: weps07test/truth_official
Answer clustering files path: weps07test/test_system
Output evaluation files path: tmp
Measures: [P, IP, FMeasure_0.5_P-IP, BEP, BER, FMeasure_0.5_BEP-BER, PM, PJ, PR,]
Baselines: [COMBINED_BASELINE, ONE_IN_ONE_BASELINE, ALL_IN_ONE_BASELINE]
Overwrite: false

Evaluating clustering answers (team truth_official) from weps07test/test_system/truth_official
Saving team evaluation to: tmp/truth_official.eval

Evaluating clustering answers (baseline COMBINED_BASELINE)
Saving team evaluation to: tmp/COMBINED_BASELINE.eval

Evaluating clustering answers (baseline ONE_IN_ONE_BASELINE)
Saving team evaluation to: tmp/ONE_IN_ONE_BASELINE.eval

Evaluating clustering answers (baseline ALL_IN_ONE_BASELINE)
Saving team evaluation to: tmp/ALL_IN_ONE_BASELINE.eval

topic	BEP	BER	FMeasure_0.5_BEP-BER	FMeasure_0.5_P-IP	IP	P	PJ	PM	PR
truth_official	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
ONE_IN_ONE_BASELINE	1,0	0,43	0,57	0,61	0,47	1,0	0,0	1,0	0,83
COMBINED_BASELINE	0,17	0,99	0,24	0,78	1,0	0,64	0,17	0,34	0,17
ALL_IN_ONE_BASELINE	0,18	0,98	0,25	0,4	1,0	0,29	0,17	0,34	0,17

Note one of the evaluation criteria is your FMeasure_0.5_BEP-PER performance; so make sure you improve the performance of your system on this measure.

We will look at your FMeasure_0.5_BEP-PER score for all names that have to be disambiguated and clustered.

The same clustering algorithm and feature set must be used for the cluster generation of all names. We will not allow and accept the usage of different clustering algorithms and/or features for different names. Practically all outputs must be generated from the same settings (clustering method and feature set)

System Output

You are expected to provide an output clustering for each person name

Abby_Watkins
Cathie_Ely
Dan_Rhone
Jane_Hunter
Michael_Howard
Thomas_Baker
Tim_Whisler

The format of the output should be the same as the Gold Standard format described above. The data output for each person name set should be created in one separate file. The file name should be the person name (blanks replaced by "_") with the ".clust.xml" extension. All these files should be put in the folder named **FirstNameStudent_LastNameStudent_CL**

Some Useful Materials

Lectures on Name Discrimination, Latent Semantic Analysis from class
http://www.isi.edu/natural-language/teaching/cs544/

SenseClusters Toolkit by Ted Pedersen:
http://www.d.umn.edu/~tpederse/senseclusters.html

Clustering in Weka:
http://weka.wikispaces.com/Using+cluster+algorithms

LSA from Infomap:
http://infomap-nlp.sourceforge.net/

LDA from Mallet:
http://mallet.cs.umass.edu/api/cc/mallet/topics/LDA.html

WebPS challenge from where data was taken:
http://nlp.uned.es/weps/weps-1/weps-1-task-guidelines

Best performing WebPS system (you can use features and tools from there):
http://acl.ldc.upenn.edu/W/W07/W07-2024.pdf