Project Text Analysis – Final Project Wikification

Malvina Nissim, Lennart Kloppenburg
m.nissim@rug.nl, lennyklb@gmail.com

27 May 2015

As we discussed in class and in the previous assignments, the final project is concerned with *wikification*. Wikification is the task of detecting entities of interest in some text, and linking them to a database. The target database that we are using is Wikipedia, and the data we are using is that from American newspapers that you have previously annotated.

In this document you will find some additional information regarding the data, the procedure, and what is required that you submit.

hint

For an overview of the task and its main components, and ideas on how to tackle issues that you most likely will have to deal with, check the **slides from Week 5**.

- use Nestor for submission
- **Deadline: Thursday 11th of June, 2015, 23:59**
- hand in:
  - your wikification system. This will have to be a python script which produces output files in exactly the same format as those you have produced manually. The only difference is the number of provided links, as you can provide up to five, if you wish. For the format of the multiple links, please see Section 1 below.
  - the output files as produced by your system, that is those that you are comparing to the gold standard.
  - your evaluation script, that is a version of `measures.py` which will produce the evaluation measures required by this project (see Section 2.2 below).
  - a report (`.pdf`) to written **individually** which contains your system's description and highlights which portions you have worked on mostly.

1 Task and Data

For Assignment 3 you had to produce some *annotated data*, including entities and Wikipedia links. As a reminder, these were the entities:

- **Country/State** — use the tag COU. All country and state names.
  Examples: France, Alaska, The Philippines, Tuscany, Burkina Faso, etc

- **City/Town** — use the tag CIT. All cities and smaller towns.
  Examples: New York, Rome, Groningen, Berlin, Zuidlaren, etc.

- **Natural places** — use the tag NAT. This includes all natural places such as lakes, mountains, volcanoes, rivers, seas, oceans, forest, etc.
  Examples: Mississippi River, Etna, Amazonia, The Pacific

- **Person** — use the tag PER. This includes all persons. Limit this to proper nouns.
  Examples: Bill Clinton, Johnny Depp, Bruce Springsteen, etc.

- **Organization** — use the tag ORG. This includes companies but also all sorts of organisations
  Examples: Google, ONU, Mercedes Benz, etc

- **Animal** — use the tag ANI. All animals.
  Examples: dog, crocodile, rabbit, cat, etc.

- **Sport** — use the tag SPO. All sports.
  Examples: football, soccer, baseball, tennis, etc.

- **Entertainment** – use the tag ENT. This includes any books, magazines, films, songs, concerts, etc.
  Examples: The Wall Street Journal, Ghostbuster, The Bible, Holes, Born to run, etc

what a file looks like is given in Figure 1. Please, note that we have added an id column at the beginning of each line, which wasn't there in the original files you got and produced.

The data that all of you have produced for Assignment 3 will be used both for developing and for testing your system, as explained below. Please, remember that you should be able to re-use all of the code that you produced for the assignments, as it was indeed intended to be useful for the final project, too.

2 Procedure

From the whole collection produced for Assignment 3, you are getting a portion which will amount to approximately 50% and which will contain all of the hand annotated information. You can use this portion to develop your system, remembering that you can always test your output against the gold standard file using the measures you already know about, to get an idea of how well you are doing. The remaining 50% will be used for testing your system, and you will only see after you will be done with developing.[1]

[1] In a standard evaluation setting, as you mentioned in class, the test file would be stripped of the information you added manually, which the evaluator would keep to produce the measures themselves. However, in this

```
ID 0 7 1001 Burkina NNP COU http://en.wikipedia.org/wiki/Burkina_Faso
ID 8 12 1002 Faso NNP COU http://en.wikipedia.org/wiki/Burkina_Faso
ID 13 14 1003 ( NNP
ID 14 22 1004 formerly RB
ID 23 28 1005 Upper NNP COU http://en.wikipedia.org/wiki/Upper_Volta
ID 29 34 1006 Volta NNP COU http://en.wikipedia.org/wiki/Upper_Volta
ID 34 35 1007 ) NNP
ID 36 44 1008 achieved VBD
ID 45 57 1009 independence NN
ID 58 62 1010 from IN
ID 63 69 1011 France NNP COU http://en.wikipedia.org/wiki/France
ID 70 72 1012 in IN
ID 73 77 1013 1960 CD
ID 77 78 1014 . .
ID 79 87 2001 Repeated NNP
ID 88 96 2002 military JJ
ID 97 102 2003 coups NNS
...
```

Figure 1: Format of gold standard file

2.1 Developing you system

The **input file** to your system should be the `.pos` file. To that, your system should add the classes and the links, wherever appropriate.

As we said in class, your system can be very basic and produce just an output file that corresponds to `en.tok.off.pos.ent`, or, on top of the production of that file (which is compulsory as we need it for evaluation), can be integrated in a web interface where a sample text can be given, and an output annotated text is returned.

hint

Demos for inspiration can be found here:

- `cogcomp.cs.illinois.edu/page/demo_view/Wikifier`
- `wikipedia-miner.cms.waikato.ac.nz/demos/annotate/`

Please remember that you system will have to output a tagged file that looks exactly like the one in Figure 1 above, with just one difference: you can produce *up to five wikipedia links* (you don't *have to*, you simply *can*). The five links should all be included in the same column, separated by a vertical pipe. Next to each link you should specify your confidence score, and it should be separated by a comma. Globally, it should look like this:

```
ID 0 7 1001 Burkina NNP COU link1,score|link2,score|link3,score|link4,score|link5,score
ID 8 12 1002 Faso NNP COU link1,score|link2,score|link3,score|link4,score|link5,score
ID 13 14 1003 ( NNP
ID 14 22 1004 formerly RB
```

case you can run the evaluation yourselves, so you will get the gold standard files, thus including the manual annotation, too. See Section 2.2 on how to use it.

```
ID 23 28 1005 Upper NNP COU link1,score|link2,score
ID 29 34 1006 Volta NNP COU link1,score|link2,score
ID 34 35 1007 ) NNP
ID 36 44 1008 achieved VBD
ID 45 57 1009 independence NN
ID 58 62 1010 from IN
ID 63 69 1011 France NNP COU link1,score|link2,score|link3,score
ID 70 72 1012 in IN
ID 73 77 1013 1960 CD
ID 77 78 1014 . .
ID 79 87 2001 Repeated NNP
ID 88 96 2002 military JJ
ID 97 102 2003 coups NNS
...
```

Remember that the order in which you list the links is important, as it will be considered as a rank (see Section 2.2 for further information on evaluation).

Once you are done, and in any case **by the deadline**, you can submit your final system and your report by uploading them on Nestor. If you upload your system before the deadline, please also send us an email to say that you're done. Also, remember that the system is jointly produced by the group, but the report has to be written **individually**. In the report you should clearly specify which portions of the project you worked on most.

2.2 Testing your system

Once you have submitted your system and the report that describes what you have done, you will receive from us the *test set*. Because you can do the evaluation yourself, the test file will also be a gold standard file thus including the entities and the link columns. On that, you can run your system and evaluate it, using precision and recall for the entities, and accuracy @K (K can be 1 to max 5) for the links. Associated to each link you will have to produce a confidence score from 0 to 1, and all scores will have to sum up to 1 for each given set of links. For example, if you provide five links, the scores could be distributed like this:

- link ranked 1st: 0.7
- link ranked 2nd: 0.1
- link ranked 3rd: 0.1
- link ranked 4th: 0.05
- link ranked 5th: 0.05

If you provide three, they could get, for example, 0.7, 0.2, and 0.1. If you provide only one, it will get a confidence score of 1. The score associated to the correct link will be your accuracy score for that link. The total accuracy will be the sum of all scores. In the example above, if the correct link was the one ranked second, you would get 0.1, if it was the one ranked first you would get 0.7. If you produce one link only and it's correct, then you would get 1. Doing it this way has the advantage of letting you at least get partial scoring instead of none, which you would get if you always had to return one link only and it wasn't correct. If you take the risk of producing always one link only and it's always correct, then you will have a total accuracy of 100%. If you always produce five links, you take less risk, but your final score will be always lower than 100% even if the correct link is always ranked first, as the total of 1 has to be distributed across the five links you have provided.

You should also produce a confusion matrix for the entities, to see what the most common mistakes were. In your presentation you will show the measures and the matrix. Please, remember that in order to compare your produced file and the gold standard file that you will get from us, including the confusion matrix, you should be able to run the same script that you developed to assess inter-annotator agreement, as the format and the measures are the same. The only addition should be the calculation of accuracy based on returning a set of up to five links.

Please, remember that at this point you shouldn't try anymore to change your system in order to get better results.

2.3 Showing your system

As a final part of this project, you will have to present your system to your fellow students and the instructors. Please, remember that each group member has to participate in the presentation. You will have to illustrate what you have done to tackle the task (what processing you've done, which tools you have used), and also present and discuss results on the test set. If you have set up a web page for interactive wikification, you are more than welcome to show a demo, too.

Presentations will will last 10–12 minutes each, and then there will be another 3–5 minutes for questions and discussion. They will take place in the week of June 16th, as announced previously, most likely in two different days. You will be required to attend both sessions, even if your group isn't presenting in one of those. The schedule will be announced soon.