Data Intensive Linguistics Assessment 2006

Philipp Koehn
Informatics
pkoehn@inf.ed.ac.uk

January 31, 2006

Named entity recognition (NER) involves the detection and classification of phrases that name entities such as persons, organisations, proteins, locations etc. It is a usefull preprocessing step for Information Extraction, Text Mining, Question Answering etc. For example, we might be interested in gathering information about company mergers from newspapers. For this an NER system would need to extract all mentions of companies within each article for further processing.

For this assignment you are to build a statistical NER system based on the material provided by the CoNLL shared task 2003. See the CoNLL webpage http://cnts.uia.ac.be/conll2003/ner/ for background details about NER and the datasets. You will also find details of a perl evaluation script which you should also use. The system will be trained and tested using the data mentioned on the CoNLL webpage. This involves two such data sets: one for English and the other for German. However,you do not need to build a NER system for both languages: one language will do.

Feel free to use any approach (in any programming language) that you think might be interesting. You can use any package or code that you want (so long as it is not a NER system!). The webpage lists a dozen or so methods, some of which are better than others.

For this assessment, you are to:

1. Develop a baseline NER system. This should be very simple and the starting point for your actual system.

2. Give a short presentation (5 minutes maximum) describing your intended approach. You should report on your baseline performance. This should be at one of the tutorials during the week starting the 13^{th} of February.

3. Develop your final NER system.

4. Write a short report (no more than 6 sides of A4) outlining the method you used, its strengths and weaknesses, and an examination of the errors made (an error analysis).

You should hand your work to Philipp Koehn (2 Buccleuch Place, 2nd floor right). It should contain the report and a printout of your code. The due date is

- 5pm on 13^{th} March 2006.

If you are going to be late submitting your work, for whatever reason, then you must tell the course secretary or course organiser (or failing that, your tutor) beforehand. Note: a standard penalty (of the awarded mark) per day operates for all late submissions. Work that is submitted more than one week late will not (usually) be accepted or marked.

Part or all of the late penalty may be waived by the exam committee in certain circumstances, such as illness, or personal problems. If you are ill please make sure that a medical certificate is submitted. Also, if other personal crises cause your work to be submitted late, please let the course secretary know.