CSCI8380 Advanced Topics in Information Systems

Spring 2010

Project 2: Entity Identification using GATE

Due: March30th

1. Take one large pile of text (i.e. from Wikipedia Documents) -- call this your *corpus*.
2. Pick a structured description of interesting things in the text (a telephone directory, or chemical taxonomy, or something from the Linked Data cloud) -- call this your *ontology*.
3. Use GATE Teamware to mark up a *gold standard* example set of annotations of the corpus (1.) relative to the ontology (2.).
4. Use GATE Developer to build a *semantic annotation pipeline* to do the annotation job automatically and measure performance against the gold standard.
5. Plot the measurements in an Excel chart.

Programming language/environment: GATE: http://gate.ac.uk/

What to submit: Please post your source code as well your data sets (corpus and ontology) and a readme file on your course web page. Also post your chart as a separate diagram. The readme file should contain: your name, a very brief info. on your implementation and other specifications you want to make.