CSC 5930/9010: Text Mining
GATE Developer Overview

Dr. Paula Matuszek
Paula.Matuszek@villanova.edu
Paula.Matuszek@gmail.com
(610) 647-9789
GATE Components

- We will deal primarily with GATE Developer:
  - It has four components:
    - Applications: groups of processes to be run on a document or corpus.
    - LanguageResources (LRs): entities such as lexicons, documents, corpora, annotation schemas, ontologies.
    - ProcessingResources (PRs): tools that operate on unstructured text, such as parsers and tokenizers. These are mostly plugins.
    - DataStores: saved processed documents and resources.
Overview of Gate Developer

- GATE Developer
- Resources Pane
  - applications: groups of processes to run on a document or corpus
  - language resources: corpus, ontologies, schemas
  - processing resources: tools that operate on unstructured text
  - datastores: saved documents and resources
- Display Pane: whatever you’re currently working with.
Language Resources

- Language Resources can be of four kinds:
  - Documents are modeled as content plus annotations plus features.
  - A Corpus is a Java Set whose members are Documents.
  - Annotations are organized in graphs, which are modeled as Java sets of Annotation.
  - Schemas are XML schemas describing allowable annotations and features.
Documents Processing in GATE

Document:

- Formats including XML, RTF, email, HTML, SGML, and plain text.
- Identified and converted into GATE annotation format.
- Processed by Processing Resources.
- Results stored in a serial data store (based on Java serialization) or indexed in a Lucene database.
- Can also be exported as XML.
New Document

• Documents are converted to GATE format; can be saved for future use or exported.

• Language Resources --> New --> Document

• Name: can leave blank and it will be created automatically (no spaces) from filename+UniqueID

• Checkmarks: required.
  • just leave defaults
  • sourceURL
    • can be a file (click the folder icon for browse)
    • or actual URL (GATE will fetch it)
    • or set to stringContent to put content in directly.

• Encoding will probably be utf-8.

• markupAware: process XML and HTML tags
Document Display

- Double-click document
  - Text (minus annotations if you chose markupAware)
  - Annotation Sets
    - from XML, HTML, previous annotation work
    - different colors for different categories
  - Annotations list
    - annotations chosen in Sets pane
Creating a Corpus

To import new documents we name the corpus and create it without any documents.

Language Resources --> New --> Corpus

Right-click and populate

- choose directory, extensions, encoding

This will create the corpus and show the corpus and the individual documents in the Resources Pane.
GATE Corpus

- Corpus Display Pane:
  - Add documents to a corpus with + button which appears when a corpus is displayed.
  - Remove with -. (Note: this removes them from corpus, not from Developer)

- Documents can be included in multiple corpora.

- A corpus can be created from a single concatenated file, by specifying the documentRootElement. This makes sense for, for instance, XML documents.
CREOLE

- A Collection of REusable Objects for Language Engineering
- The set of resources integrated with GATE
- All the resources are packaged as Java Archive (or ‘JAR’) files, plus some XML configuration data.
- Managed in the Creole Plugin Manager
Processing Resources: ANNIE

- A family of Processing Resources for language analysis included with GATE
- Stands for A Nearly-New Information Extraction system.
- Using finite state techniques to implement various tasks: tokenization, semantic tagging, verb phrase chunking, and so on.
ANNIE IE Modules

http://gate.ac.uk/sale/tao/splitch6.html#chap:annie
Some ANNIE Components

- Tokenizer
- Gazetteer: lists of entities
- Sentence Splitter
- Part of Speech Tagger
  - produces a part-of-speech tag as an annotation on each word or symbol.
- Semantic Tagger
ANNIE Component: Tokenizer

- Token Types
  - word, number, symbol, punctuation, and space

- A tokenizer rule has a left hand side and a right hand side.
Example Tokenizer Rule

"UPPERCASE_LETTER" "LOWERCASE_LETTER"*

> 

Token;orth=upperInitial;kind=word;

- The sequence must begin with an uppercase letter, followed by zero or more lowercase letters. This sequence will then be annotated as type “Token”. The attribute “orth” (orthography) has the value “upperInitial”; the attribute “kind” has the value “word”.
ANNIE Component: Gazetteer

• The gazetteer lists used are plain text files, with one entry per line.
• Each list represents a set of names, such as names of cities, organizations, days of the week, etc.
Example Gazetteer List

• A small section of the list for units of currency:
  
  • ...... 
  
  • Ecu
  European Currency Units
  FFr
  Fr
  German mark
  German marks
  New Taiwan dollar
  New Taiwan dollars
  NT dollar
  NT dollars
  
  • ......
ANNIE Component: Sentence Splitter

- Segments the text into sentences.
- This module is required for the tagger.
- The splitter uses a gazetteer list of abbreviations to help distinguish sentence-marking full stops from other kinds.
ANNIE Component: Semantic Tagger

- Based on JAPE language, which contains rules that act on annotations assigned in earlier phases.
- Produce outputs of annotated entities.
Example Using ANNIE

- http://services.gate.ac.uk/annie/

- More next week.
Create an Application with Processing Resources (PRs)

- Applications model a control strategy for the execution of PRs.
- Simple pipelines: group a set of PRs together in order and execute them in turn.
- Corpus pipelines: open each document in the corpus in turn, set that document as a runtime parameter on each PR, run all the PRs on the corpus, then close the document.
- We will do a simple application during lab.
LOTS more

- GATE is an extraordinarily rich system. Some other CREOLE resources included in the standard distribution:
  - Annotation Merging, Quality assurance summarizer for comparing annotations
  - Web crawler, Information Retrieval, Key Phrase Extraction
  - Machine learning
  - Domain-specific taggers (e.g., chemistry)
  - Resources for many languages
  - CREOLE plugins for integrating with many other systems.

- More details at http://gate.ac.uk/gate/doc/plugins.html
Some Links

- Home page is [http://gate.ac.uk/](http://gate.ac.uk/)
- Some good short tutorial videos for getting started: [http://gate.ac.uk/demos/developer-videos/](http://gate.ac.uk/demos/developer-videos/) . These are only a few minutes each, so they’re fast. Version 6, but they don’t seem to be very different.
- Lots of documentation (“acres” of it): [http://gate.ac.uk/documentation.html](http://gate.ac.uk/documentation.html)
- The wiki: [http://gate.ac.uk/wiki/](http://gate.ac.uk/wiki/)
- Some very nice course materials, with a lot more detail than we will cover: [http://gate.ac.uk/wiki/training-materials-2011.html](http://gate.ac.uk/wiki/training-materials-2011.html)
What Next?

• In lab we will create a simple application and use it.
• Next week we will go into a lot more detail on using Annie for information extraction
• Homework. (You knew that was coming...)
• We will not cover using the APIs for GATE or most of the advanced applications. This might be the best tool for some of your projects, though.