Introduction to GATE

Majid Sazvar
sazvar@stu-mail.um.ac.ir

Knowledge Engineering Research Group
&
Web Technology Laboratory
Ferdowsi University of Mashhad

2010

What is GATE? 1/2

- GATE (General Architecture for Text Engineering) is a free, open source software platform for Natural Language Processing.
- Originally developed at the University of Sheffield beginning in 1995.

The University Of Sheffield.

- GATE excels at text analysis of all shapes and sizes. From large corporations to small startups, from multi-million research consortia to undergraduate projects.

What is GATE? 2/2

- GATE is the biggest open source language processing project with a development team more than double the size of the largest comparable projects (many of which are integrated with GATE):
  - LingPipe,
  - OpenNLP,
  - UIMA,
  - and many more specific tools.

- GATE Homepage:

http://gate.ac.uk/

GATE Facilities 1/2

- GATE is:
  - an IDE, **GATE Developer**: an integrated development environment for language processing components bundled with a very widely used Information Extraction system and a comprehensive set of other plugins.
  - a web app, **GATE Teamware**: a collaborative annotation environment for factory-style semantic annotation projects built around a workflow engine and a heavily-optimized backend service infrastructure.
  - a framework, **GATE Embedded**: an object library optimized for inclusion in diverse applications giving access to all the services used by GATE Developer and more.

GATE Facilities 2/2

- In the future, GATE will have:
  - a wiki/CMS, **GATE Wiki** (http://gatewiki.sf.net/), mainly to host our own websites and as a testbed for some of our experiments.
  - a cloud computing solution for hosted large-scale text processing, **GATE Cloud** (http://gatecloud.net/).

What we can do with GATE? 1/2

- GATE includes components for diverse language processing tasks, e.g. parsers, morphology, tagging, Information Retrieval tools, Information Extraction components for various languages, and many others.
- Visualization and editing of annotations, ontologies, parse trees, etc.
- A finite state transduction language for rapid prototyping and efficient implementation of shallow analysis methods (JAPE)
- Measurement, evaluation, benchmarking (never believe a computing researcher who hasn't measured their results in a repeatable and open setting!)
- Pluggable machine learning implementations (Weka, SVM Light, ...)

What we can do with GATE? 2/2

- GATE Developer and Embedded are supplied with an Information Extraction system (ANNIE) which has been adapted and evaluated very widely (numerous industrial systems, research systems evaluated in MUC, TREC, ACE, DUC, Pascal, NTCIR, etc.).

- ANNIE is often used to create RDF or OWL (metadata) for unstructured content (semantic annotation).

GATE Architecture 1/2

- GATE as an architecture suggests that the elements of software systems that process natural language can usefully be broken down into various types of component, known as resources.

- GATE components are specialized types of Java Bean, and come in three flavors:
  - Language Resources (LRs) represent entities such as lexicons, corpora or ontologies;
  - Processing Resources (PRs) represent entities that are primarily algorithmic, such as parsers, generators or ngram modelers;
  - Visual Resources (VRs) represent visualization and editing components that participate in GUIs.

GATE Architecture 2/2

- Collectively, the set of resources integrated with GATE is known as CREOLE: a Collection of REusable Objects for Language Engineering.

- All the resources are packaged as Java Archive (or `JAR‘) files, plus some XML configuration data. The JAR and XML files are made available to GATE by putting them on a web server, or simply placing them in the local file space.

GATE Plugins 1/4

- Alignment
- ANNIE
- Annotation_Merging
- Copy_Annots_Between_Docs
- Gazetteer_LKB
- Gazetteer_Ontology_Based
- Groovy
- Information_Retrieval
- Inter_Annotator_Agreement
- Jape_Compiler
- Keyphrase_Extraction_Algorithm
- Language_Identification

GATE Plugins 2/4

- Lang_Arabic
- Lang_Cebuano
- Lang_Chinese
- Lang_Hindi
- Lang_Romanian
- Learning
- LingPipe
- Machine_Learning
- Ontology
- Ontology_BDM_Computation
- Ontology_OWLIM2
- Ontology_Tools

GATE Plugins 3/4

- OpenNLP
- Parser_Minipar
- Parser_RASP
- Parser_Stanford
- Parser_SUPPLE
- Schema_Annotation_Editor
- Stemmer_Snowball
- Tagger_Abner
- Tagger_Chemistry
- Tagger_Framework
- Tagger_MetaMap
- Tagger_NP_Chunking
- Tagger_OpenCalais

GATE Plugins 4/4

- Tools
- UIMA
- Web_Crawler_Websphinx
- Web_Search_Google
- Web_Search_Yahoo
- Web_Translate_Google
- WordNet

Downloading GATE

- The latest stable version of GATE is Release 6.0 (November 8th 2010).
- You can download it from:

http://gate.ac.uk/download/

Further Reading

- Lots of documentation lives on the GATE web server, including:
  - Movies of the system in operation;
  - The main system documentation tree;
  - JavaDoc API documentation;
  - HTML of the source code;
  - Parts of the requirements analysis that version 3 was based on.

?