An open source solution for full lifecycle Text Analytics

GATE

General Architecture for Text Engineering
http://gate.ac.uk/

The University Of Sheffield.

Free

Open source, licensed under LGPL allowing unrestricted commercial use, hosted on SourceForge.

100% Java

Runs on **any platform** supporting Java 5 or later. Developed and tested daily on Linux, Windows, and Mac OS X.

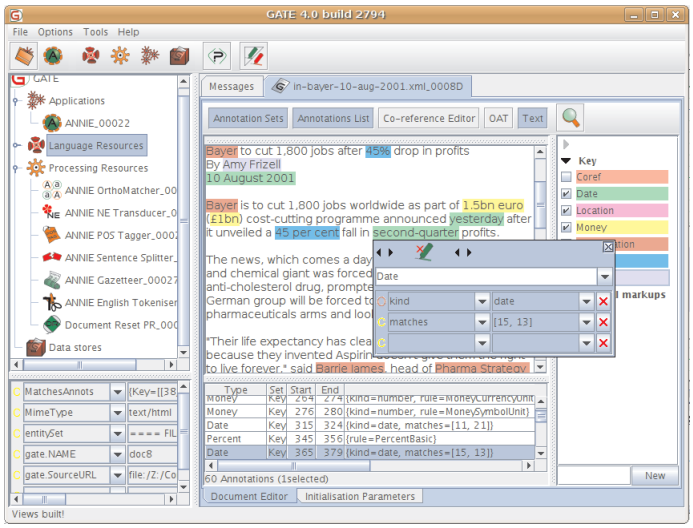
Mature and actively supported

In development **since 1996**; now at version 5.0; around 20 active developers.

Comprehensive

Support for manual annotation, performance evaluation, information extraction, [semi-]automatic semantic annotation, and many other tasks.

Over **50 plugins** included with the standard distribution, containing over 70 resource types. Many others available from independent sources.



Standards-based

Reference implementation in **ISO** TC37/SC4 LIRICS project; supports XCES, ACE, TREC etc. formats; founder member of **OASIS**/UIMA committee.

Efficient

Optimisations included with the latest version provide a 20 to 40% speed and memory usage improvement.

Highly efficient finite state text processing engine; many plugins with linear execution time.

Popular

Assessed as "outstanding" and "internationally leading" by an anonymous EPSRC peer review.

Used at thousands of sites: companies, universities and research laboratories, all over the world. ~35,000 downloads/year.

Rolling funding for more than 15 staff at the University of Sheffield.

Data Management

Pluggable input filters with out of the box support for XML, HTML, PDF, MS Word, email, plain text, etc.

Common in-memory data model built around stand-off annotation, documents and corpora.

Persistent storage layer with support for XML or Java serialisation. I/O interoperation with many other systems.

Standard Algorithms

Ready made implementations for many typical NLP tasks such as tokenisation, POS tagging, sentence splitting, named entity recognition, co-reference resolution, machine learning, etc.

User Interface

Comprehensive tool set for data editing and visualisation, rapid application development, manual annotation, ontology management.

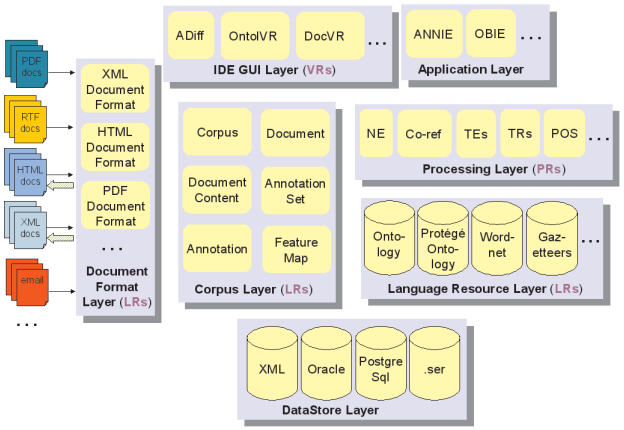
Integration

Leveraging the power of other projects such as:

- **Information Retrieval:** Lucene (Nutch, Solr), Google and Yahoo search APIs, MG4J;
- **Machine Learning:** Weka, MaxEnt, SVMLight, etc.;
- **Ontology Support:** Sesame and OWLIM;
- **Parsing:** RASP, Minipar, and SUPPLE;
- **Other:** UIMA, Wordnet, Snowball, etc.

Community and Support

Friendly and active community of developers and users offers efficient help. Commercial support available from Ontotext and Matrixware.



Overview

GATE was first released in 1996, then completely re-designed, re-written, and re-released in 2002. The system is now one of the most widely-used systems of its type and is a comprehensive infrastructure for language processing software development.

The new UIMA architecture from IBM/Apache has taken inspiration from GATE and IBM have paid the University of Sheffield to develop an interoperability layer between the two systems.

Key features of GATE are:

- Component-based development reduces the systems integration overhead in collaborative research.
- Automatic performance measurement of Language Engineering (LE) components promotes quantitative comparative evaluation.
- Distinction between low-level tasks such as data storage, data visualisation, discovery and loading of components and the high-level language processing tasks.
- Clean separation between between data structures and algorithms that process human language.
- Consistent use of standard mechanisms for components to communicate data about language, and use of open standards such as Unicode and XML.
- Insulation from idiosyncratic data formats (GATE performs automatic format conversion and enables uniform access to linguistic data).
- Provision of a baseline set of LE components that can be extended and/or replaced by users as required.

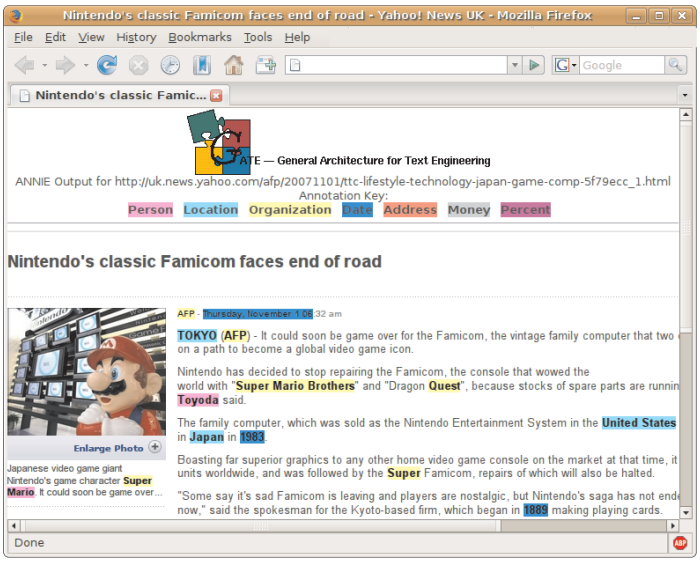
Text Analysis

Text Analysis (TA) is a process which takes unseen texts as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in Information Retrieval (IR) applications.

TA covers a family of applications including named entity recognition, relation extraction, event detection.

GATE has been used for **TA applications** in domains including bioinformatics, health and safety, and 17th century court reports.

TA systems built on GATE have been evaluated among the top ones at **international competitions** (MUC, ACE, Pascal). A system built by the GATE team came top in two of three categories in the NTCIR 2007 patent classification competition.



The GATE family

- GATE Developer: an integrated development environment for language processing components bundled with the most widely used Information Extraction system and a comprehensive set of other plugins
- GATE Embedded: an object library optimised for inclusion in diverse applications giving access to all the services used by GATE Developer and more
- GATE Teamware a collaborative annotation environment for high volume factory-style semantic annotation projects built around a workflow engine and the GATE Cloud backend web services
- GATE Cloud: a parallel distributed processing engine that combines GATE Embedded with a heavily optimised service infrastructure

First cousins: the Ontotext family

- Ontotext KIM: UIs demonstrating our multiparadigm approach to information management, navigation and search
- Ontotext Mímir: (Multi-paradigm Information Management Index and Repository) a massively scaleable multiparadigm index built on Ontotext's semantic repository family, GATE's annotation structures database plus full-text indexing from MG4J

Sponsored by: Ontotext.com, Matrixware.com
Research funding: EU, UK Research Councils and JISC

Contact: Prof. Hamish Cunningham
http://www.dcs.shef.ac.uk/~hamish/