Infrastructure for Human Language Technology

GATE

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 newer. Developed and tested daily on Linux, Windows, Mac OS X, and Solaris.

Mature and actively supported

In development for **over 12 years**; current project version 4.0; around 20 active developers.

Comprehensive

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

Over **30 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.

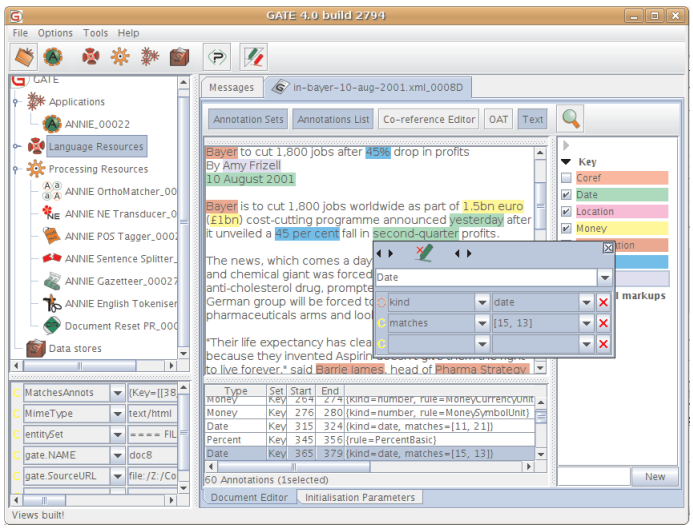
Highly efficient finite state text processing engine; many plugins with linear execution time.

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. Over **20,000 downloads** in the last 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, Oracle, PostgreSQL, 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.

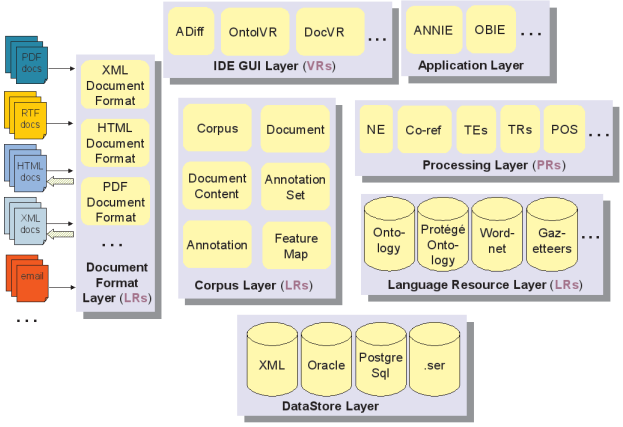
Integration

Leveraging the power of other projects such as:

- **Information Retrieval:** Lucene, Google and Yahoo search APIs;
- **Machine Learning:** Weka and SVMLight;
- **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 also available.



Overview

GATE, a General Architecture for Text Engineering, 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 relatively 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.

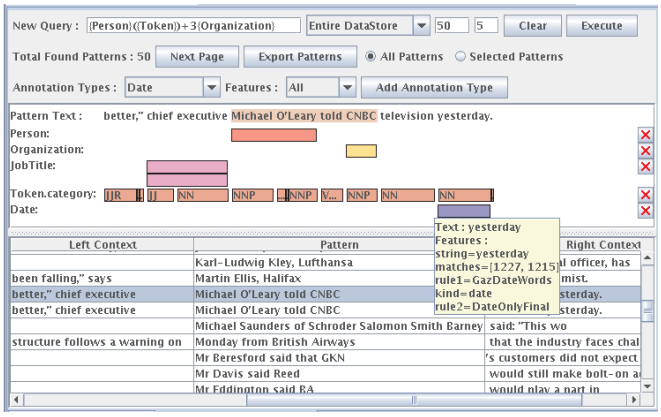
Information Extraction

Information Extraction (IE) 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.

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

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

IE 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.



IE Development Tools

The standard GATE distribution includes **ANNIE**, an Information Extraction system that can be used an example or a starting point for customisations. Example code is provided for embedding ANNIE into other applications.

Knowledge Engineering

GATE has facilities for finite state processing over annotations based on regular expressions:

- **JAPE** is a pattern-action language, where linguistic patterns are matched and annotations are created as a result.
- **ANNIC** (pictured left) is a visual tool for assisted rule development.

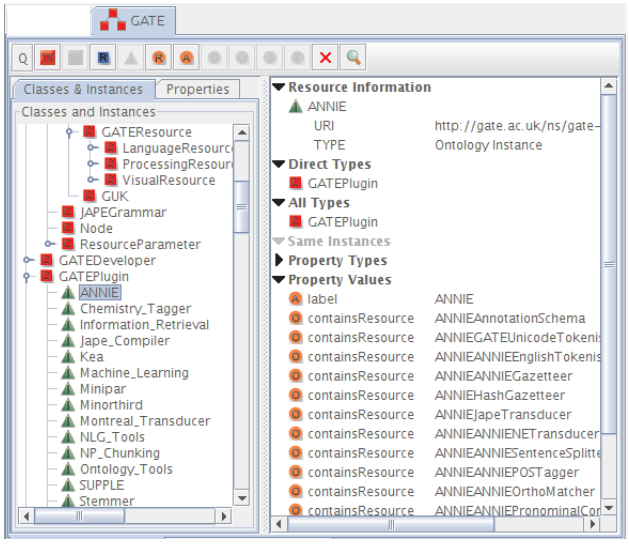
Machine Learning

In aid of developers for IE systems, GATE includes support for Machine Learning for text mining, entity recognition, and relation extraction. An abstraction layer unifies access to different ML algorithms.

Working with Ontologies

Ontology Abstraction Layer

Based on Sesame RDF store (http://openrdf.org), with additional OWL support provided by OWLIM (http://www.ontotext.com/owlim/), leading to one of the fastest and most scalable triple stores. Ontologies can be loaded with storage in memory, on disk or on a dedicated server.



Ontologies in GATE

Taxonomical relations can be used in annotation matching, thus enhancing JAPE's power of generalisation.

Graphic interface tools for ontology visualisation, ontology editing, and semantic annotation of text are included with GATE.

Ontology Learning

Automatically extending ontologies with knowledge extracted from text through Information Extraction.

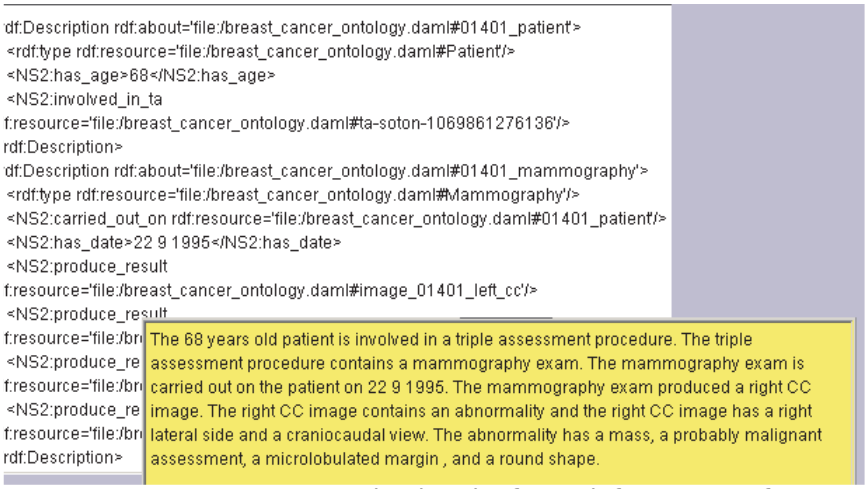
Knowledge Base Population

Automatically populating knowledge bases with instance data extracted from text. This is related to Semantic Annotation.

An Example

The GATE ontology shown to the right has been automatically derived from source code and associated meta-data and automatically populated from the software and user documentation.

Research Projects



E-Science

- **AKT:** Advanced Knowledge Technologies;
- **MultiFlora:** support for biodiversity research;
- **MiAKT:** collaborative problem solving in medical informatics;
- **CLEF:** tools for integrating patient information from text and images.

Digital Libraries

- **GATE/ETCSL:** automatic morphological analysis of literary Sumerian texts;
- **EMILLE:** building a 63M words corpus of South Asian languages;
- **OldBaileyIE:** named entity recognition on 17th century Old Bailey court reports;
- **PrestoSpace:** automatic creation of meta-data for news broadcasts used for advanced indexing and conceptual search.

Semantic Web and knowledge Technologies

- **SEKT:** Next Generation Knowledge Management;
- **hTechSight:** building a KM platform for the chemical industry;
- **TAO:** migrating legacy applications to open, semantics-based, service oriented architectures;
- **NeON:** shaping the future infrastructure for semantic applications;
- **Musing**: knowledge extraction from multiple sources and ontology population for business intelligence applications;
- **MediaCampaign:** automating the detection and tracking of media campaigns on television, Internet and in the press .

Human Language Technology

- **MUSE:** Named entity recognition from diverse text types;
- **SAFE:** collaborative mixed initiative semantic annotation;
- **LIRICS:** definition of an ISO standard for language technology with a reference implementation.

Semantic Annotation

Automatic and semi-automatic production of **semantic meta-data** for text and multimedia. GATE identifies mentions of known concepts and instances from an ontology. This type of meta-data enables a search by meaning paradigm to enhance traditional retrieval methods. Searches like "find companies located in Western Europe involved in the high tech sector" become possible.

RichNews

RichNews is an example of applying semantic annotation to **multimedia broadcasts**. It combines timing information from automatic speech recognition and semantics extracted from news web pages. The meta-data produced is then used for advanced retrieval facilities.

Science

GATE is a platform for experimental repeatability, quantitative evaluation, collaborative development and integration.

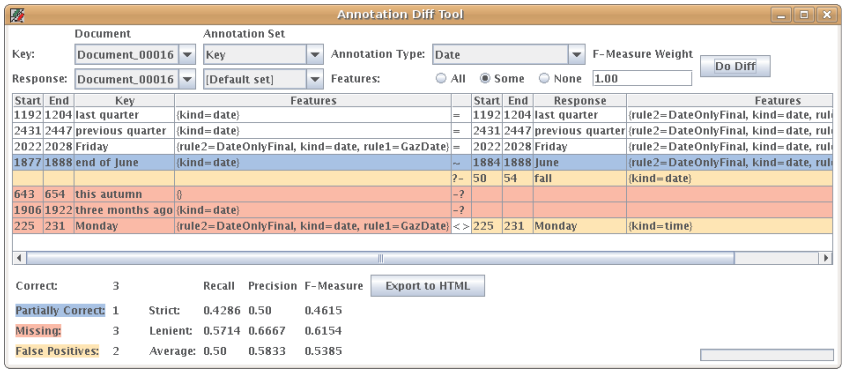
Education

GATE and the ANNIE IE system are used in classrooms across the world and in many postgraduate projects.

Business

GATE has been engineered to a high standard in order to be suitable for deployment in commercial applications software, and is based on components, mobile code and internet-based distribution. A serious effort has been made to achieve a very high level of quality; unit and regression tests are run nightly on three different platforms.

Our IE software is quality-controlled and Sheffield has applied IE in very many domains, and developed World-leading expertise in producing robust systems for diverse applications.



Commercial users include: Glaxo Smith Kline PLC, AT&T, Master Foods NV, British Gas PLC, Syntalex Ltd., Lernout Hauspie Gmbh., Thompson Corp, Innovantage, Garlik, Fizzback, Spock.

Acknowledgements

GATE research is or has been funded by the EU, EPSRC, AHRB, BBSRC, and commercial grants.

Contact:

Prof. Hamish Cunningham

hamish@dcs.shef.ac.uk