The GATE Family for text processing

http://gate.ac.uk

GATE MÍMIR · GATE DEVELOPER · GATE EMBEDDED · GATE CLOUD · GATE TEAMWARE

GATE Family

- GATE is a family of tools for natural language processing.
- It includes support for a wide range of tasks, such as named entity resolution, opinion mining, semantic annotation, in a variety of languages.
- It has been under active development since 1997, at the University of Sheffield and other contributing sites.
- The current implementation uses the Java platform, which allows it to be deployed on any operating system, including Windows, Mac OS, or Linux.

Open Source

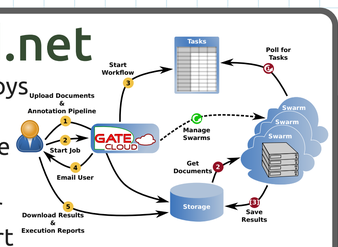
GATE software is open source, under the LGPL licence, which encourages commercial use.

Big Data

Commodity hardware and cloud computing have made big data approaches possible for small and medium organisations. However, a significant fraction of today's data comes in the form of unstructured text, which cannot be used directly. GATE helps identify the underlying structure of text by performing linguistic and semantic analysis.

GATECloud.net

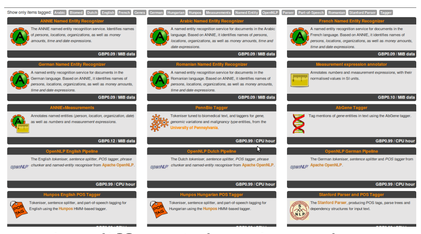
- GATECloud.net deploys GATE processes in a Software-as-a-Service architecture.
- Any user can register with the site and start processing their first million documents within minutes.
- Users can upload their own analysis pipelines, or can use one of the ready-made ones.
- Mímir and Teamware servers can also be rented; usage is charged by the hour.



AnnoMarket.com

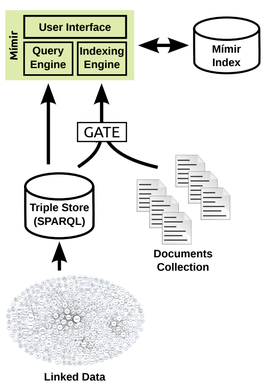
AnnoMarket.com extends the facilities of GATECloud.net with:

- An API for remote programmatic access to the cloud platform.
- A large number of ready-built pipelines, covering different domains and many languages.
- Support for automatic production of document sets via custom web crawling.



GATE Mímir

- Mímir is a framework that provides integrated semantic search over document text, annotations, and semantics.
- It has been used to index tens of millions of documents and provide advanced search facilities for collections including archives of government web sites, patent documents, and scientific literature.



Social Media

- Much of new text is now user-generated through social media. This type of content can be very important in settings such as politics, customer relationship, or trend spotting.
- Processing and understanding text from social media is hard: messages are often brief, containing mistakes, and don't provide background.
- Several GATE tools have been produced that deal with social media and user-generated content.

TwitIE

A GATE application for extracting information from tweets, one of the noisiest forms of social media text. It performs:

- language identification;
- custom tokenisation, that deals with user names, smileys, #-tags, URLs, etc.;
- part-of-speech tagging, using models trained to work on Twitter text;
- text normalisation;
- named entity recognition.

Opinion Mining

Of particular interest when working with social media text is the ability to detect the opinions and sentiments expressed. GATE has tools to do that, which are also able to detect some cases of sarcasm.

Opinion	Text
very positive	Navigating in #Hadoop filesystem with basic unix command is really awesome!
negative	Spent half an hour fiddling with the digital radio because @BBCRadio4 was broadcasting a cooking progamme rather than @BBCr4today #unhappy
positive	ROME: Ever the unpredictable showman, **Silvio Berlusconi** used a special televised address on Wednesday to stake a claim for a continued place at the heart of Italian politics.
negative	I hope we get to hear even more about the 2 English players who DIDN'T make the Lions squad.

Intelligent Archives

Archives are increasingly concerned with providing intelligent access to their content. GATE modules target the semantic dimension of archival material, and enable semantic preservation and access by:

- extracting semantic information such as named entities, terminology, events, opinions from web resources;
- enriching named entities with links to the LOD cloud;
- providing metadata for archival objects in the form of populated ontologies.

The National Archives (UK)

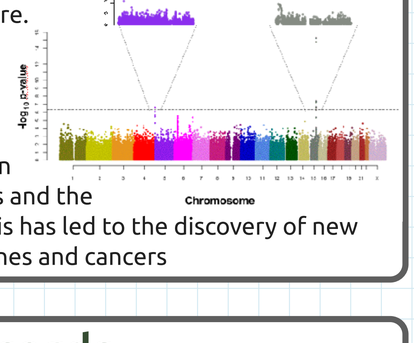
In a project with The National Archives (UK), a GATE analysis pipeline was used to process the archived versions of all the government web sites. It identified generic entities such person, organisation, and location names, date and time expressions, amounts of money. Also recognised were domain-specific entities, such as the names of government departments and official positions. The annotated pages were then indexed using Mímir, and the resulting index was used to power new semantic-aware search modalities over the archive contents.

Life Sciences

- Life sciences research is one of the areas most affected by information overload. PubMed already includes more than 20,000,000 records and keeps growing ever faster.
- GATE helps by automatically extracting information from scientific literature and other documents.

Cancer Research

In the LarKC project, GATE tools were used to systematically process scientific literature. The output was a set of prior probabilities that were plugged into a statistical model estimating the correlation between genetic markers and the occurrence of cancer. This has led to the discovery of new associations between genes and cancers

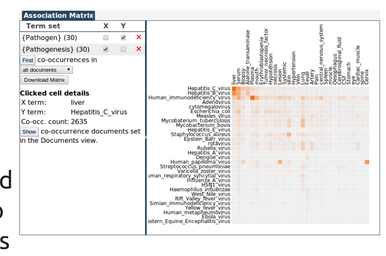


Hospital Records

- A large amount of patient information is recorded in the free text of patient records, and not in the structured part of those records. This makes it difficult to access.
- GATE has been used to extract diagnoses, medications, symptoms and tests from patient records, for use by medical researchers, and in day to day patient care.
- GATE has also been used in several commercial electronic patient record systems.

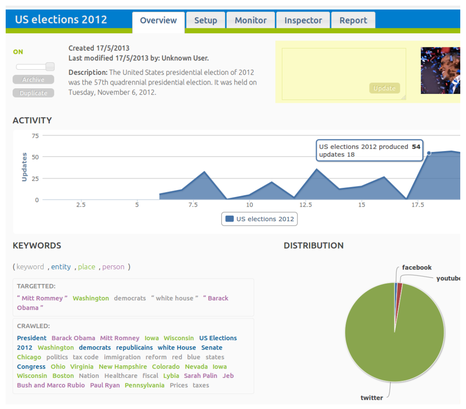
Analytics

Semantic analysis of scientific literature, combined with the integrated semantic search facilities provided by Mímir, can be used to power complex analytics tools. This screen-shot uses the depth of colour to show the association strength between pathogens and entities involved in pathogenesis, such as organs.



ARCOMEM

- ARCOMEM (http://www.arcomem.eu) creates innovative content appraisal and selection processes for web archiving and preservation, via an adaptive decision support tool for the archivist.
- GATE tools analyse textual content from social media and other web sources, performing entity, event and opinion extraction.
- This semantic analysis adds meaningful, socially aware contextualisation to the archived content, enabling categorisation and reasoning, and assisting journalists to answer the big 5 questions (who, what, when, where, why).



EPSRC Engineering and Physical Sciences Research Council
SEVENTH FRAMEWORK PROGRAMME
BBSRC
Arts & Humanities Research Council

GATE has been supported by and has contributed to the outputs of several research projects, funded under grants by the European Union, the UK Engineering and Physical Sciences Research Council, the UK Arts and Humanities Research Council, the UK Biotechnology and Biological Sciences Research Council, and others.

GATE is an open-source project under the stewardship of the University of Sheffield.

The University Of Sheffield.