

GATE - a General Architecture for Text Engineering <http://gate.ac.uk>

Kalina Bontcheva, Hamish Cunningham

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK
kalina,hamish@dcs.shef.ac.uk

GATE¹ is an architecture, development environment, and framework for building systems that process human language (Cunningham et al., 2002; Maynard et al., 2002b). It has been in development at the University of Sheffield since 1995, and has been used for many R&D projects (Maynard et al., 2000), including Information Extraction in multiple languages and media, and for multiple tasks and clients. GATE is available freely, as an open source system, under the GNU library licence. Version 2 has been completely redeveloped in Java, and is a stable, robust, and scalable infrastructure for Human Language Technology, which allows users to focus on language processing issues, while mundane tasks like data storage, format analysis and data visualisation are handled by GATE.

The new version 2, demonstrated here, differs significantly from the previous version, not only because it is capable of processing larger texts and is faster and more robust; other features are:

- JAPE - a rule-based language which provides finite state transduction over annotations based on regular expressions;
- comprehensive multilingual support via additions to Java's Unicode capabilities;
- performance evaluation tools;
- corpus markup support;
- reusability of visualisation resources, as well as language and processing resources;

¹This work has been supported by the Engineering and Physical Sciences Research Council (EPSRC) under grants GR/K25267 and GR/M31699, and by several smaller grants.

- database support for storing language resources (Oracle, PostgreSQL);
- support for distributed resources from the Web;
- improved robustness, scalability, and efficiency;
- more comprehensive document format support: XML, SGML, HTML, e-mail, RTF, plain text.

A family of Processing Resources for language analysis is also included in the shape of ANNIE, A Nearly-New Information Extraction system. The majority of these components use JAPE-based finite state techniques to implement various tasks from tokenisation to semantic tagging and coreference. The emphasis is on robustness and low-overhead portability, rather than full parsing and deep semantic analysis. The set of currently distributed modules comprises: tokeniser, gazetteer, sentence splitter, part-of-speech tagger, named entity recognition grammars, coreference resolution, and chunking. We are currently working on adding learning algorithms, e.g., Hidden Markov Models.

While all of these modules are English-specific, our recent experience has shown that some can be reused directly (e.g., the tokeniser can handle Indo-European languages) and/or easily customised for new languages (Pastra et al., 2002). The ANNIE modules typically also separate clearly the linguistic data from the algorithms that use it, thus allowing modules to be adapted to new domains/languages by just modifying the data itself.

Another strand of recent GATE developments to be demonstrated is oriented towards supporting HLT applications for the Semantic Web. GATE now handles *ontologies* as a new type of language resource, which can be accessed and manipulated by any language processing module, including the JAPE-based ones. Ontologies can be created and edited visually via the popular Protégé editor (Noy et al., 2001), which has been integrated with GATE's visual application-development environment.

Another new feature is the *Information Retrieval* support, which allows for applications that can benefit from combining HLT and IR. The implementation is based on the freely-available Lucene system².

The demonstration will also show some applications built using GATE and its freely available modules:

- MUMIS: *MUltiMedia Indexing and Search* - uses GATE-based Information Extraction (IE) components as an off-line annotation provider via XML with a multimedia search tool. The IE modules identify metadata, entities, and events (Saggion et al., 2002).
- *Named-entity based summarization* of company reports: this is a real-world application, where the task has previously been performed manually by humans. The aim of the system is to summarise information from the reports in order to generate statistics about the level of compliance with Health and Safety recommendations and legislation (Maynard et al., 2002a).
- *Robust cross-domain named entity recognition* for (semi-)automatic Semantic Web annotation.

References

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th*

Anniversary Meeting of the Association for Computational Linguistics.

- D. Maynard, H. Cunningham, K. Bontcheva, R. Catizone, George Demetriou, Robert Gaizauskas, Oana Hamza, Mark Hepple, Patrick Herring, Brian Mitchell, Michael Oakes, Wim Peters, Andrea Setzer, Mark Stevenson, Valentin Tablan, Christian Ursu, and Yorick Wilks. 2000. A Survey of Uses of GATE. Technical Report CS-00-06, Department of Computer Science, University of Sheffield.
- D. Maynard, K. Bontcheva, H. Saggion, H. Cunningham, and O. Hamza. 2002a. Using a text engineering framework to build an extendable and portable ie-based summarisation system. In *Proceedings of the ACL Workshop on Text Summarization*. forthcoming.
- D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. 2002b. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering - Special Issue on Robust Methods in Analysis of Natural Language Data*. forthcoming.
- N.F. Noy, M. Sintek, S. Decker, M. Crubzy, R.W. Ferguson, and M.A. Musen. 2001. Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems*, 16(2):60-71.
- K. Pastra, D. Maynard, H. Cunningham, O. Hamza, and Y. Wilks. 2002. How feasible is the reuse of grammars for Named Entity Recognition? In *Proceedings of 3rd Language Resources and Evaluation Conference*. forthcoming.
- H. Saggion, H. Cunningham, D. Maynard, K. Bontcheva, O. Hamza, C. Ursu, and Y. Wilks. 2002. Extracting Information for Automatic Indexing of Multimedia Material. In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, page xxx, Las Palmas, Gran Canaria, Spain.

²See <http://jakarta.apache.org/lucene/docs/index.html>