GATE

Infrastructure for Text Mining and Natural Language Processing

http://gate.ac.uk

CONTACT:
Prof: Hamish Cunningham
hamish@dcs.shef.ac.uk

DATA MANAGEMENT
- Pluggable input format filters
- Common data model of stand-off annotation
- Persistent storage: XML, database, etc.

INTEGRATION
- **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.

STANDARDS-BASED
- ISO TC37/SC4; founder member of OASIS/UIMA
- XCES, ACE, TREC formats

EFFICIENT
- 20–40% speed and memory improvements
- Many components with linear performance

POPULAR
- "Outstanding" and "Internationally leading"
- Used at thousands of sites;
- Over 20,000 downloads in the last year

FREE
— Open source, LGPL, on SourceForge

MATURE AND ACTIVELY SUPPORTED
— Over 12 yrs. old; version 4.0
— Community support: over 20 active developers

COMPREHENSIVE
— Over **30 plugins** & 70 components

MANY NLP ALGORITHMS
— Named Entity Recognition
— Information Extraction
— Semantic Annotation
— Text and Web Mining
— Document Summarisation
— Performance Evaluation

100% JAVA
— Platform-indepedent
— Tested on Linux, Windows, Mac OS X
— Java 5 or newer

Information Extraction: Development Support

GATE includes ANNIE, an Information Extraction system that can be used an example or a starting point for customisations

KNOWLEDGE ENGINEERING

Finite state processing over annotations based on regular expressions:
- **JAPE** is a pattern-action language for matching linguistic data
  - **ANNIC:** visual tool for assisted rule development

MACHINE LEARNING
- Machine Learning framework for text mining, entity recognition, and relation extraction
  - An abstraction layer for unified access to different ML algorithms

PERFORMANCE EVALUATION TOOLS

Applications

SEMANTIC ANNOTATION
- (Semi-)automatic production of **semantic meta-data** for text and multimedia
- Identify mentions of known concepts and instances from an ontology
- Enables a search by meaning paradigm to enhance traditional retrieval methods

RICHNEWS
- Applying semantic annotation to **multimedia broadcasts**
- Combines timing information from video and semantics from web pages

SELECTED COMMERCIAL USERS
- Glaxo Smith Kline PLC
- AT&T
- Master Foods NV
- British Gas PLC
- Syntalex Ltd.
- Lernout Hauspie Gmbh.
- Thompson Corp
- Innovantage
- Garlik
- Fizzback
- Spock

RESEARCH PROJECTS
- **E-Science:** AKT, MultiFlora, MiAKT, CLEF
- **Digital Libraries:** Prestospace, GATE/ETCSL, EMILLE, OldBaileyIE
- **Semantic Web and Knowledge Management:** SEKT, TAO, NeON, MediaCampaign
- **Human Language Technology:** SAFE, LIRICS, MUSE