

The University of Sheffield Computer Science Department http://www.dcs.shef.ac.uk



The Sheffield NLP group http://nlp.shef.ac.uk



E — General Architecture for Text Engineering

http://gate.ac.uk/

GATE is an architecture, framework and development environment for R&D in language and knowledge technologies. The system is free software under the GNU Library licence, and includes:

• stable, robust and scaleable language processing tools;

•integration with Semantic Web infrastructure such as Protégé, Sesame and KAON and support for XML, RDF(S), DAML+OIL, etc.;

• components for Information Retrieval, annotation, evaluation, visualisation, persistence, treebanking, enhanced Unicode support, machine learning; Digital Libraries, etc. etc.

GATE and its NLE components have been successfully used in a large number of research projects and commercial applications, including Information Extraction in multiple languages, from multimedia sources, and for multiple tasks and clients. The system has been in development at the University of Sheffield since 1995. A partial snapshot of current GATE users includes:

- Pearson Educational, UK;
- Merck KgAa, Germany;
- •Canon Europe, UK;
- Knight Ridder, US;
- •BBN Technologies, US;
- Reuters Inc., US;
- OntoText / Sirma AI Ltd., Bulgaria;
- Resco AB, Sweden/Finland/Germany;
- Glaxo Smith Kline PLC, UK;
- the American National Corpus project, US;
 the Perseus Digital Library project,

• Imperial College, London, the University of Manchester, Queen Mary College, UMIST, the University of Karlsruhe, Vassar College, ISI (the University of Southern California) and a large number of other Universities around the world.

• the Perseus Digital Library project Tufts University, US;

Education <u>Http://gate.ac.uk/teaching.html</u>

Postgraduates in locations as diverse as Sofia, Copenhagen and London are using the system in order to avoid having to write simple things like sentence splitters from scratch, and to enable visualisation and management of data. For example, <u>Partha Lal</u> at Imperial College is developing a summarisation system based on GATE and ANNIE. (His site includes the URL of his components; give GATE the URL and it will load his software over the network.) <u>Marin</u> <u>Dimitrov</u> of the University of Sofia has produced an <u>anaphora resolution system</u> for GATE.

Scientific Research Http://gate.ac.uk/science.html

- **Repeatability**: making it easier to repeat comparable experiments across different sites.
- Quantitative evaluation: built-in quantitative metrics, e.g. precision and recall.
- Collaboration: easy software integration and porting.
- **Reuse not reinvention**: reusing results does not require learning fresh installation and usage conventions for every tool.

Commercial Applications

http://gate.ac.uk/business.html

- Engineered to a high standard for deployment in commercial applications software.
- Based on components, mobile code and internetbased distribution.
- •100% Java with support for XML, HTML and relational databases (Oracle, PostgreSQL).
- •Large cross-platform regression test suite.
- IE software quality-controlled by the rigorous application of **quantitative evaluation** metrics that ensure that the behaviour of our systems is **predictable** and **robust.**

GATE has been funded by the EPSRC http://www.epsrc.ac.uk/

Contact:

Prof. Yorick Wilks, or Dr. Hamish Cunningham Senior Research Scientist Department of Computer Science University of Sheffield, UK. Email: hamish@dcs.shef.ac.uk Web: http://gate.ac.uk/hamish Phone: +44 114 222 1891

Portable Information Extraction

http://gate.ac.uk/ie/

GATE is distributed with a **Portable Information Extraction** component set called ANNIE, useable in many different applications:

- Copes seamlessly with documents in many different formats, from badly-spelled lower case email messages to structured XML or HTML pages to newswires to 18th century court reports.
- Able to process large data volumes without crashing and at high speed, scaling from desktops to very mainframes and clusters.
- System developers can adapt the system to new circumstances with a minimum of effort.
- Users can adapt the system as far as is possible (some IE tasks cannot be attempted by unskilled users, but where the data is simple end-users can and should be allowed to update the system).
- ANNIE is in use in Romanian, Bulgarian, Greek, Bengali, Spanish, Swedish, German, Italian, French (Arabic, Chinese and Russian next year).
- Integrated with the WEKA machine learning toolkit and the OntoText HMM suite.

MUltiMedia Indexing and Searching environment



Digital Libraries

Http://gate.ac.uk/digilibs.html/

As digital libraries grow in size and coverage, so does the need for automatic content annotation and indexing. GATE's robust and customisable Named Entity recognition and Information Extraction technology has already been used successfully for metadata creation, automatic name and event annotation, indexing, and access. Applications:

- OldBaileyIE required adaption to the nonstandard written conventions of Old English in Old Bailey court reports from the 18th Century;
- in MUMIS (Multimedia Indexing and Search) we annotate material in multiple modalities to build a conceptual index of football videos;
- EMILLE focuses on collection and annotation of large corpora in non-indigenous minority languages in the UK (Urdu, Bengali, Sylheti and others).

We are currently looking at GATE for the creation of computational tools for the study of digital collections in cultural heritage languages, e.g. Ancient Greek, Latin.

Multilingual Language Resources Http://gate.ac.uk/tao/

- Facilities for developing annotated corpora and other Language Resources (LRs).
- Annotation model compatible with the XCES and ATLAS systems; Xschema type model.
- Visualisation and editing tools support trees, chains and flat annotations structures.
- Fully supports multilingual LRs with Unicode facilities beyond default Java:
 - a Unicode editor with input methods for many languages;
 - use of the input methods in all places where text is edited in the GUI;
 - a development kit for implementing input methods;
 - ability to read diverse character encodings.

🐵 Gate Unicode Editor - Unicode Sampler.txt 🛛 😨 🗖 🔀
File Edit Options Help
😅 🔛 😰 🗠 😹 📭 🌇 Arial Unicode MS 💌 18 💌
.أنا قادر على أكل الزجاح و هذا لا بؤلمني :Arabic
Armenian: Կընամ ապակի ուտել և ինձի անհանգիստ չըներ
Chinese: 我能吞下玻璃而不伤身体。
من مي توانم بدون احساس درد شيشه بخورم. :Farsi / Persian
Georgian: მინას ვჭამ და არა მტკივა.
Hebrew: אני יכול לאכול זכוכית וזה לא מזיק לי.
Hindi: मैं काँच खा सकता है, मुझे उस से कोई पीडा नहीं होती.
Japanese:私はガラスを食べられます。それは私を傷つけません。
Korean: 나는 유리를 먹을 수 있어요. 그래도 아프지 않아요
Marathi: मी काच खाऊ शकतो, मला ते दुखत नाही.
ره شیشه خورلې شم، هغه ما نه خوروی :Pashto
Sanskrit: कार्च शक्रोम्यतुम् । नोपिहनिस्त माम् ।
Thai: ฉันกินกระจกได้ แต่มันไม่ทำให้ฉันเจ็บ
میں کانچ کھا سکتا ہوں اور مجھے تکلیف نہیں ہوتی ۔ :Urdu
Yiddish: איך קען עסן גלאָז און עס טוט מיר נישט וויי.

The Semantic Web

Http://gate.ac.uk/semweb.html

The Semantic Web is adding a machine-tractable layer to the natural language web of HTML. The benefits of success will be many, but the project is currently lacking the critical mass necessary to demonstrate these benefits beyond a few small-scale trial applications. GATE is being used for experiments in automatic and semi-automatic methods for:

•linking web pages to Ontologies using Information Extraction;

•learning and evolving Ontologies via IE and lexical semantic network traversal.

We have also integrated the Protégé Ontology editor with the system. GATE forms the basis of the language technology under development in the UK's <u>Advanced Knowledge Technologies</u> six-year multi-site programme.

Text Annotations Annotation Sets Coreference Print	29 E	
By Schehorstrade Danestikhu and Andrew Parker, Ort 09 2004	no no no	
by othereracade panesining and platest carbon of the	ConTEXT - [C:\TEMP\It-coalition-leaders-08-oct-2001.htm_00066daml.xml]	
Within minutes of the first missile strikes against Afghanistan	Elle Edit View Format Project Tools Options Window Help	
began voicing their support for Fresident George W. Bush. Bu	B C C C C C C C C C C C C C C C C C C C	Docu
a 'terrorist attack' which would 'never achieve its goar',	The coalition leaders 08 oct 2001 htm 00065dani und	
The US president telephoned a number of coalition leaders in g	Anthermony Assessed Suchashing Transmission (
president and Jacques Chirac, president of France.		
	annitoncology cariabout	
Mr.Chirat said military operations would take place over a long	demi:versioninro* 1.0 >	1.5.7
Infrastructure of the terrorist networks," he said. Alam Richard.	<pre><dami:comment>autogenerated from GATE R0FFormatExporte</dami:comment></pre>	EE </td
active partner in the offensive "in a matter of days".	<pre>countioncology></pre>	
A Russian Foreign Ministry statement said Afghanistan had bu	<pre> (gate:Organization rdf:about='European Union'/></pre>	
a safe haven for "terrorists" responsible for crimes in many co-	<gate:location rdf:about="Pakistan"></gate:location>	
said	<pre><gate:location rdf:sbout="France"></gate:location></pre>	
	<pre><gate:person rdf:about="Vladimir Putin"></gate:person></pre>	
Some 40 countries were co-operating, Mr Bush said. This incl	<pre><gate:person rdf:about="Gerhard Schröder"></gate:person></pre>	
	<gate:location rdf:about~'america'=""></gate:location>	
Germany's chancellor, Gerhard-Schrüder, said it supported the	<gate:location rdf:about="Niddle East"></gate:location>	
	<gate:person rdf:sbout="Mr Bush"></gate:person>	
Belgian Prime Minister Guy Vemoticad, whose country curren	<pre> daml:sameIndividual&s rdf:resource='Bush'</pre>	
the presidency reanimied solidanty with the US.	rdf:type= http://pillango.sirma.bg/daml/news.daml	L#Pe
Tow Dire 116 prime minister, declared that the US and Eriters		
rause is his? The nime minister justified military article and	<pre> {gate:Ferson rdf:about='Tony Blair'></pre>	
over Osama bin Laden and members of his a-Questa terroris a	<pre></pre>	
	rdf:type='http://pillango.sirma.bg/daml/news.daml	1/Pe
"This cause is just," said Mr Blan, "The murder of almost 7,001		
way of life, and civilised values the world over.* The British par	<pre><gate:person rdf:about="Hr Chirac"></gate:person></pre>	
session since the September 11 attacks.	<pre>daml:sameIndividual&s rdf:resource*'Jacques Chirac'</pre>	
	cdf:type='http://pillangu.sirma.bg/daml/news.daml	L#Fe
Britain has made available to the US its airbase on Diego Oat		-
		owned in case of the local division of the l



Information Retrieval Http://gate.ac.uk/tao/

- Full-featured Information Retrieval for queries on corpora.
- Documents can be retrieved textual content plus annotations.
- Current implementation based on Lucene; others may be added in future.

🔉 Gate	Messages 🗚 Corput Searcher 💉 Search Application		
Applications Search Application Language Resources ATE corpus_0003D Processing Resources A Corpus Searcher Data stores Tile/D/temp/ds/	Document	Score	
	gu-ecb-10-aug-2001.xml_000D0	0.237	
	ft-SSL-10-aug-2001.xmi_0008E	0.22	
	ft-bt-03-aug-2001.xml_00055	0.186	
	ft-Industrial-gloom-07-Aug-2001.xml_00082	0.166	
	ft-bt-at&t-01-jul-2001 xml_0005E	0.179	
	gu-telewest-10-aug-2001.xmi_00114	0.179	
	ft-bt-26-jul-2001.xml_0005B	0.155	
	ft-WestLB-BT-05-aug-2001.xml_00094	0.155	
	R-bank-of-england-02-aug-2001.xml_00046	0.143	
	gu-ECB-07-aug-2001.xml_000CD	0.143	
	gu-ECB-03-aug-2001 xml_000CA	0.131	

Dialogue Http://gate.ac.uk/dialogue.html/

GATE is being used in the Amities project to produce dialogue processing server components to run in the Galaxy Communicator architecture, which concentrates on distributed processing, hooking together sets of servers and clients that collaborate to hold dialogues with human interlocutors. GATE ignores the issue of distributed processing, and concentrates on facilities for bootstrapping, developing, testing and deploying language processing components. We used GATE to produce a Galaxy Communicator server component, and there seems a natural synergy between the two systems. In future work we would like to more closely integrate GATE with Galaxy Communicator.

Evaluation Http://gate.ac.uk/sale/tao/

Two mechanisms for automated performance measurement and visualisation of the results.

- •Annotation Diff compares and shows differences on a single document, and produces Precision, Recall, F-measure and Error Rate statistics.
- •Regression Test tool tracks system over time over a whole corpus.

Ker

260

264

2894

752

23/5

2636

933

Select the KEY annotation Messages GATE corpus_0003D ANNIE_0001E example document Default set Text Annotations Annotation Sets Coreference Print Ø, String - ResponseStern Fayland 2353 UK 255 - KeyEnd - Key 2365 Coreference data "It's an indication that Mike Armstrong is serious about local competition and serious about getting moving," said Anna-Marie Kovacs, an analyst for the E Default brokerage firm Janney Montgomery Scott Inc. "AT&T and Teleport are going after ATaT Europe 746 the business market, which is where local companies make a lot of their money." Wales IIK 231 V Telepor 2801 The deal also will give three major cable television companies, which are majority owners of Teleport, a collective 10 percent stake in AT&T UK 931 By acquiring Teleport, AT&T can offer business customers local and Recall strict: 0.5557 Recall average: 0.5557 Recall lenient: 0.5557 long-distance esterda telephone service, and data and Internet access, under its own brand name Using 's local facilities, the company also would be able to reduce the fees WorldCom Inc it pays to local phone companies for access to local telephone customers. Mike Armstri "It's going to permit us to be much more cost-effective as we go for that local business," Armstrong said at a news briefing. "This has competition and growth written all over it." AT&T is paying for Teleport with its stock. Teleport shareholders will receive 0.943 AT&T shares for each of their Teleport shares, putting the deal at \$ 59 a share based on AT&T's closing price vesterday of \$ 82.62 1/2 a share, up \$ 2.62 1/2 Teleport closed down \$ 3.82 1/2 at \$ 54.12 1/2 a share. The commanies expect the deal, which must be annowed by regulators Annotations Editor Features Editor

GATE has been developed by:

\Xi Arnola in 2775 il

String - Keystari Fogland 2358

Hampshire 2631

258

2886

2370

2801

2628

931

Precision strict: 1.0000 Precision average: 1.0000 Precision leptent: 1.0000

746

It airlines

It airlines

Swanwich

Swanwick

Curope

Wales

IIK.

UK.

UK

lect the KCY dat

Hamish Cunningham, Valentin Tablan, Kalina Bontcheva, Diana Maynard, Marin Dimitrov,

Cristian Ursu, Oana Hamza, Bobby Popov, Angel Kirilov, Atanas Kiryakov,

Damyan Ognyanov, Horacio Saggion, Mark Leisher, Wim Peters

Coreference Viewer