Institution: University of Sheffield

Unit of Assessment: 11 - Computer Science and Informatics

Title of case study: GATE: General Architecture for Text Engineering

1. Summary of the impact

GATE (a General Architecture for Text Engineering—see http://gate.ac.uk/) is an experimental apparatus, R&D platform and software suite with very wide impact in society and industry. There are many examples of applications: the UK National Archive uses it to provide sophisticated search mechanisms over its .gov.uk holdings; Oracle includes it in its semantics offering; Garlik Ltd. uses it to mine the web for data that might lead to identity theft; Innovantage uses it in intelligent recruiting products; Fizzback uses it for customer feedback analysis; the British Library uses it for environmental science literature indexing; the Stationery Office for value-added services on top of their legal databases. It has been adopted as a fundamental piece of web infrastructure by major organisations like the BBC, Euromoney and the Press Association, enabling them to integrate huge volumes of data with up-to-the-minute currency at an affordable cost, delivering cost savings and new products.

2. Underpinning research

In the 1990s The University of Sheffield's Department of Computer Science founded a research programme, led for the past 15 years by Professor Cunningham (with initial EPSRC grants to Professors Wilks and Gaizauskas, GR/K25267/01 and GR/M31699/01, begun in 1994 and 1999 respectively), in text analysis and the software architectures appropriate to its efficient development, measurement, adaptation, deployment and maintenance [R1, R2]. The programme has grown to be a world leader, attracting some £12m direct funding from RCUK, EC and industry.

Our research aim was to provide a general software framework for text engineering. The first stage of the research was to conduct an analysis of the text processing field, with a particular focus on the software architectures, methodological approaches and infrastructural elements present in available systems. This research enabled us to develop a core software framework for text engineering including: the appropriate object classes for flexibly and efficiently modelling text processing components; the best compromise between expressive power and efficiency for pattern matching over textual annotations; the requirements for component frameworks supporting extensibility of the architecture; effective ways to allow combination of statistical counts with derivations based on linguistic intuition.

We tested this framework through successive iterations of an open source code base, which was also deployed in further RCUK research (funded by EPSRC, BBSRC, and AHRC) and applied in collaborative EU-funded projects and for industrial customers. We identified common patterns in the development processes that typified successful technology transfer projects in our field (and some common human factors in these processes). We developed a process specification and a set of support tools that encapsulate these patterns and promote repeatability of successful results in new projects.

The GATE framework currently comprises: **GATE Developer** an integrated development environment (IDE) for language processing components, which is bundled with several hundred plugins; **GATE Cloud**, a cloud computing solution for hosted big data text processing; **GATE Teamware**, a collaborative environment for large-scale manual semantic annotation projects, a multi-paradigm index server; **GATE Mimir**, which can be used to index and search over text, annotations, semantic schemas (ontologies), and semantic meta-data (instances), allowing queries that arbitrarily mix full-text, structural, linguistic and semantic constraints and that can scale to terabytes of text; and the **GATE Embedded** object library that has been optimized for inclusion in diverse applications giving access to all the services mentioned above.

Core GATE ideas and design are described in 3 key papers [R1, R2, R3], which together have over 2000 citations. Further refinements and applications of GATE are described in PLoS One [R6], PLoS Computational Biology [R5], Philosophical Transactions of the Royal Society A [R4], and the Journal of Web Semantics. There are three books about the system. With popular annual summer schools and huge number of downloads (details in next section), GATE has, as envisaged by the initial research programme, become the standard open source resource for Natural

Language Engineering for both research and commercial users.

GATE comes with a suite of state-of-the-art methods for **extracting information** from the web, news wires, scientific journal papers, and legal and medical documents. The GATE IE components use both rule-based and machine learning methods. These were developed as part of the EPSRC Advanced Knowledge Technologies project (GR/N15764) and the SEKT EU project. Our SVM-based IE system was among the best evaluated in the PASCAL challenge on machine learning for IE, and in the NTCIR patent classification task. GATE was also the first text mining platform to support Ontology-Based Information Extraction (OBIE) from RDF/OWL ontologies [R3] and more recently multi-billion instance Linked Open Data ontological resources. Cunningham has received close to £1 million in industrial funding to develop robust, scalable IE from patent databases, and in 2011-12 was ANR Chaire d'Excellence at the Internet Memory Foundation in Paris applying our methods to **multi-terabyte web crawls**. Through GATE Bontcheva is developing new IE and text summarisation methods, suited to the short, noisy, colloquial, and highly contextual nature of **social media** (EPSRC Fellowship EP/I004327/1 and the TrendMiner project).

Most recently, we developed a **novel and unique cloud-based platform for large-scale NLP research** – http://gatecloud.net (JISC/EPSRC grant EP/I034092/1). It gives NLP and other researchers (e.g. digital humanities, eHealth, eScience) access to NLP algorithms and enables them to carry out large-scale NLP experiments by harnessing the vast, compute power of the Amazon cloud. It also reduces the need to implement specialized parallelisable text processing algorithms [R4]. Further research on adapting GATE to language processing over big data on the cloud is being carried out within the EC-funded AnnoMarket project, coordinated by Cunningham.

3. References to the research *(** denotes outputs which best demonstrate underpinning research quality)*

R1. **H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proc. of the 40th Anniv. Meeting of the Association for Computational Linguistics (ACL'02). 2002. [1443 citations]

R2. H. Cunningham. *GATE, A General Architecture for Text Engineering*. Computers and the Humanities, volume 36, pp. 223-254, 2002. doi: 10.1023/A:1014348124664 [451 citations]

R3. K. Bontcheva, V. Tablan, D. Maynard, H. Cunningham. *Evolving GATE to Meet New Challenges in Language Engineering*. Journal of Natural Language Engineering. 10 (3/4), pp. 349-373. 2004. doi: 10.1017/S1351324904003468 [179 citations]

R4. **V. Tablan, I. Roberts, H. Cunningham, K. Bontcheva. *GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud*. Philosophical Transactions of the Royal Society A, 371(1983), 2013 doi: 10.1098/rsta.2012.0071.

R5. **H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. *Getting More out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics*. PLoS Computational Biology. 9(2): e1002854, 2013, doi: 10.1371/journal.pcbi.1002854.

R6. M. Johansson, A. Roberts, D. Chen, Y. Li,..., G. Byrnes, H. Cunningham, P. Brennan, J. Wakefield, and J.D.Mckay. Using Prior Information from the Medical Literature in GWAS of Oral Cancer Identifies Novel Susceptibility Variant on Chromosome 4 -- the AdAPT Method. PLoS ONE 7(5):e36888. May 2012, doi: 10.1371/journal.pone.0036888.

4. Details of the impact

Based only on collaborations directly involving us, GATE's text mining algorithms have been used by companies to build products targeting diverse verticals including: media (delivery; analysis; journalism); pharmaceuticals; patent search; voice-of-the-customer; brand, product, and reputation management; social media analytics; bioinformatics. Its major commercial beneficiaries have been both large organisations (BT, Elsevier, Yahoo, Atos, Dassault Aviation, MPS Bank, Creditreform, BBC, the Press Association, Euromoney) and SMEs and start-ups (Text Mining Solutions, Intellius, Fizzback, Garlik, Innovantage, Ontotext, Mondeca, Playence, ELDA, CoreSystems).

A 2010 study of the market for intelligent text processing solutions (Grimes, [S5]) estimated its size at $835m, with 25-40% growth potential per annum. Growth was driven by the technology's central role in social media analysis and by text analytics' contribution to advanced semantic search and search-based applications. The study lists GATE as one of the leading technologies in the sector.

There are 54 US patents which reference the GATE framework, including 18 from IBM, 11 individual patents, other patents from Xerox, AT&T, Hewlett Packard, BT, and Research in Motion. This reflects the fact that GATE enables the wide take-up of text processing technology.

Since 2010 the GATE Summer Schools, funded by commercial fee-paying attendees (who make up around half the audience) have been attended by 210 people representing 74 distinct commercial organisations including Adobe, Bayer, Elsevier, the Health Protection Agency and Lockheed Martin. The software has had over 190,000 downloads since 2008 [S9].

GATE has had very substantial economic impact. It has allowed the foundation of new businesses and existing businesses to exploit new markets and make cost savings on existing products. We provide exemplars in *Digital Semantic Publishing* and *Semantic Search and Information Extraction*.

Digital Semantic Publishing**:** The **BBC**'s GATE-based digital semantic publishing platform has achieved savings of approximately 80% in the editorial costs of running the news and sports website compared to a conventional database-backed web system. These systems represent capital investment of millions of pounds. The platform is integrated right across the BBC News and Sports sites, and has made "*a huge improvement to the quality of the site that wouldn't have been possible without GATE*" according to the Chief Technical Officer at the Financial Times, formerly Director of Architecture and Development at the Press Association and Chief Architect of BBC News and Sports interactive, who has integrated GATE in all three organisations (S1). "*The breadth and speed of BBC online coverage of major events like the World Cup and the Olympics 2012 would have been technologically and financially impossible without it.*" The cost savings arise from the platform's ability to allow journalists to develop and change content for web publishing without involving teams of technologists. Consider a London Olympics example: in a conventional approach, if a relatively unknown athlete achieved a gold medal, increased interest would warrant a new area on the site dedicated to that athlete. This would require adjustments to the relational database structure, tailoring of search indices and consultation on the content prominence between technical staff and journalists. Digital semantic publishing uses GATE to connect the journalistic content to a *semantic* repository, automating this process. The creation of the content becomes simply a matter of journalist choice, rather than requiring broad consultation.

The **Press Association** "*have adopted GATE as part of their core infrastructure for the faster and better production of the Wire and all the services they offer, but it has also allowed them to as develop a new offering, supplying tailored content to specific customers*" [S1]. For example, GATE made it possible for them to offer an affordable high quality solution to link up meta data and push all content to the London Committee of the Olympic Games (LOCOG) website. Before GATE, it would have been too difficult and costly to offer such a service at a competitive price and still make a profit. The site was a worldwide hit, the biggest in the world (excluding Google and other search engines) in terms of content and every possible usage metric for the duration of the games [S1].

Other organisations are embracing the technology**. CNN** commissioned a similar system in 2011, it is already part of the web infrastructure for **Euromoney**, and a GATE-based platform is currently in development by the **Financial Times** for the better management and categorisation of information**. "***It is now one of the key tools in the web development toolset*" [S1].

Semantic Search and Information Extraction**: Fizzback** processes customer feedback for major UK transportation clients like First and National Express. Their information extraction architecture is based on GATE. Fizzback was founded in 2004, received £1.9 million second round investment in 2009, and was sold to **Nice Systems** for USD80 million in 2011 [S6]. Operating in the *voice of the customer* market, Fizzback provides a service that would have been impossible without the GATE language analysis tools on which it was based, according to its founding CTO [S2]. **Foodity,** co-founded in 2010 by the founding CTO of Fizzback, processes recipes from the web using GATE to derive their ingredient lists. They secured £300k in mid-2012 bringing their total funding to £450k. Their entire business model would not have been possible without the predictability, maturity and wide spectrum tool support of GATE. "As an early adopter of GATE I have never looked back. My notable successes have been with Fizzback and Foodity both as shining examples of NLP integration in an enterprise application" [S2].

Ontotext is one of the EU's largest independent semantic technology businesses with 70 employees, and turnover of €3m in 2012. The company's entire technology stack is based on GATE's text mining functionality. Executive Director Atanas Kiryakov confirms: "*The use of GATE*

in our products is considered very beneficial by our clients, because it guarantees them low vendor lock-in… There were several cases in which Ontotext won projects against other text mining vendors because of the openness and popularity of GATE…Overall since 2008, Ontotext have secured contracts for more than €5m on work that directly uses GATE" (S3). Examples of GATE-based projects include: a joint venture targeting the UK recruitment industry called **Innovantage** (2008) that harvests job and vacancies data from over 30,000 company websites and 300 job boards to create currently the richest database of UK job offerings; and enabling the **UK National Archives** to index and provide a semantic search for over 10TB of government web pages.

Text Mining Solutions is a startup formed in April 2011 that began delivering information extraction solutions in January 2013. The company generated turnover of £8k in its first three months of trading, and is on track to generate forecast turnover of £50k in its first full financial year. Company Director Stephen Brewer confirms: "*It was my awareness of GATE and the commercial possibilities it offered that initially led me to found Text Mining Solutions, and the technology is absolutely fundamental to our whole service offering. Our company, and the two full time posts it has created, simply would not exist without GATE*" [S4]. The company has five key customers, with some twenty more in prospect. GATE is used both to mine text on behalf of some customers and to write rules for others to integrate into their own GATE pipelines. "*The key advantage that GATE offers ... is its combination of flexibility with scaleability. It is modular and so can be adapted to a hugely diverse range of individual customer needs, but it can also be scaled up for implementation on large projects handling significant volumes of data … The UK is leading the way in the adoption of text mining technology, and GATE is making a major contribution internationally to knowledge and information management within an emerging market that wouldn't be possible without it. This is a high-potential market that is still in the early adopter phase, and GATE has allowed Text Mining Solutions to position itself securely ahead of competitors*" [S4].

IBM have been "inspired and influenced" by GATE [S7] leading them to adopt new technologies. Specifically, GATE has had a critical impact on the development of their Unstructured Information Management (UIMA) software. UIMA underpins IBM's content analytics offering and their Watson question answering system, key parts of what they see as a $20bn opportunity in Business Intelligence and Analytics (IBM CIO interview [S8]). In 2001 IBM began work on their UIMA system "a software architecture for defining and composing interoperable text and multi-modal analytics". UIMA interoperates with GATE through a translation layer that connects the two systems allowing UIMA users access to GATE analytics, a capability that IBM deemed sufficiently important to directly fund Cunningham to develop in 2005. IBM have now released UIMA, including the GATE interoperability layer, under Apache license (uima.apache.org). **IBM** have 18 awarded patents referencing GATE.

5. Sources to corroborate the impact

S1. Current CTO of the Financial Times; previously Director of Architecture and Development, the Press Association; Chief Architect, BBC News and Sport Interactive can confirm impact on these three organisations.

S2. Co-founder and director at Foodity; previously: CTO at FizzBack Group.

S3. Letter from CEO of Ontotext corroborates the impact of GATE on their business model.

S4. Email from Director and Founder of Text Mining Solutions Ltd confirms importance of GATE.

S5. http://www.b-eye-network.com/view/15219 corroborates market size and GATE as leading technology.

S6. http://www.nice.com/nice-acquire-fizzback-introducing-most-complete-customer-experience-management-offering-integration

S7. David Ferrucci, et al. (2006). Towards an Interoperability Standard for Text and Multi-Modal Analytics. IBM Research Report, RC24122 (W0611-188), page 4.

S8. http://practicalanalytics.wordpress.com/2011/11/02/ibm-cio-study-bi-and-analytics-are-1-priority-for-2012/

S9. http://sourceforge.net/projects/gate/files/stats/timeline?dates=2006-01-01+to+2013-09-24