Institution: University of Sussex

Unit of Assessment: UoA 11 Informatics

Title of case study:

Automatic grammatical analysis enabling advanced text processing in commercial applications

1. Summary of the impact

Research carried out at Sussex into the automatic grammatical analysis of English text has enabled and enhanced a range of commercial text-processing applications and services. These include an automatic SMS question-answering service and a computer system that grades essays written by learners of English as a second language. Over the REF period there has been substantial economic impact on a spin-out company, whose viability has been established through revenue of around £500k from licensing, development and maintenance contracts for these applications.

2. Underpinning research

The research in automatic grammatical analysis (parsing) of English text underpinning this case study was led at Sussex by John Carroll, in collaboration with a team of academic and industrial researchers at Sussex and elsewhere. At Sussex, Carroll was a Research Fellow, 1996–2001, Reader, 2001–6, and Professor from 2006. The work was partially funded by the following EPSRC grants and EU contracts for which Carroll was the PI:

- Robust Analysis of Unrestricted English Text (EPSRC GR/A00751 and GR/L02135, 1996–2001, total £176,185). Building on Carroll's previous work on the Alvey Natural Language Tools (ANLT), over the course of this Advanced Fellowship grant he played a key role in creating:
  - the first parser based on a manually-developed, linguistically-informed grammar that was able to deal effectively with unrestricted natural language text, and which covers a wide range of both formal and informal constructions of English – in contrast to parsers induced from syntactically annotated text *corpora* (or 'treebanks'), which are typically restricted to a single language genre;
  - an approach to representing and processing language ambiguity efficiently by means of subsumption operations over grammatical categories, allowing the parser to process millions of words of text in little time; and
  - the first demonstration of improvements to parsing accuracy through the use of a non-deterministic language-processing pipeline, allowing ambiguity resolution to be postponed to later processing stages without degrading overall system performance [see Section 3, R1, R4].
- Shallow Parsing and Knowledge Extraction for Language Engineering (SPARKLE) (EU FP4 LE12111 subcontract, 1996–7, £64,615). Pioneering contributions included:
  - a representation for parser output based on grammatical relations between words, which is suitable for interfacing to natural language application systems; and
  - a technique for automatically inferring knowledge about the grammatical behaviour of words from large amounts of unannotated text, allowing a parser to be tuned to a specific domain without the expensive requirement for further syntactically annotated text [R2].
- PSET: Practical Simplification of English Text (EPSRC GR/L53175, 1998–2001, £167,783)

Significant novel research included:

- an efficient, reversible lemmatiser for English based on industry-standard finite-state tools;
- an accurate method for guessing the likely part of speech of a previously unseen word, allowing a parser to process text containing spelling mistakes, abbreviations, acronyms and rare or technical vocabulary; and
- methods for evaluating parsers via grammatical-relation representations, making it possible to reliably compare different parsers and to precisely focus development efforts [R3, R4].

- Robust Accurate Statistical Parsing (RASP) (EPSRC GR/N36493, 2001–4, £170,414) Innovative work included:
  - The robust recovery of fragmentary analyses from ungrammatical text and non-standard language usage, based on computing probabilities for partial parses; and
  - methods for computing a set of probabilistically-weighted grammatical relations, as a faithful and computationally-tractable representation of the full space of possible analyses for a sentence [R4, R5].

These research contributions are all implemented in the RASP system for the automatic grammatical analysis of English text. The first version of RASP was released in 2002; since 2011, RASP has been available as open source (http://ilexir.co.uk/applications/rasp/) under the GNU Lesser GPL.

3. References to the research

R1 Carroll, J. and Briscoe, E. (1996) 'Apportioning development effort in a probabilistic LR parsing system through evaluation', in Brill, E. and Church, K. (eds) *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadephia: University of Pennsylvania, 92–100, http://aclweb.org/anthology/W96-0209 (60 citations, Google Scholar).

R2 Carroll, J., Minnen, G. and Briscoe, E. (1998) 'Can subcategorisation probabilities help a statistical parser?', in Charniak, E. (ed.) *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*. Montreal: University of Montreal, 118–126. http://aclweb.org/anthology/W98-1114 (76 citations, Google Scholar)

R3 Minnen, G., Carroll, J. and Pearce, D. (2001) 'Applied morphological processing of English', *Natural Language Engineering*, 7(3): 225–250, doi: 10.1017/S1351324901002728 (241 citations, Google Scholar).

R4 Briscoe, E. and Carroll, J. (2002) 'Robust accurate statistical annotation of general text', in González Rodríguez, M. and Suárez Araujo, C.P. (eds) *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*. Las Palmas: University of Las Palmas in Gran Canaria, 1499–1504. http://www.lrec-conf.org/proceedings/lrec2002/pdf/250.pdf (299 citations, Google Scholar)

R5 Briscoe, E., Carroll, J. and Watson, R. (2006) 'The second release of the RASP system', in *Proceedings of the ACL-COLING'06 Interactive Presentation Sessions*. Sydney: University of Sydney, 77–80, http://aclweb.org/anthology/P06-4020 (273 citations, Google Scholar).

R6 Andersen, O., Nioche, J., Briscoe, E. and Carroll, J. (2008) 'The BNC parsed with RASP4UIMA', in *Proceedings of the Sixth Language Resources and Evaluation Conference (LREC)*. Marrakech, Morocco, 865–869. http://www.lrec-conf.org/proceedings/lrec2008/pdf/218_paper.pdf (21 citations, Google Scholar).

In total there are almost 1,000 citations of these outputs, indicating substantial academic significance. Evidence for the originality and rigour of outputs [R1], [R3] and [R5] comes from the fact that they are published in well-established journals and highly selective international

conferences.

Outputs [R1], [R3] and [R5] best indicate the quality of the underpinning research at a level that is at least internationally recognised.

Outputs can be supplied by the HEI on request.

4. Details of the impact

In 2003, Carroll (jointly with Ted Briscoe of Cambridge University) founded a spin-out company – iLexIR Ltd – to commercially exploit the research summarised above. Since 2008, the company has developed an extended version of RASP incorporating features that ease the task of integrating it into large-scale application systems. These include:

- adding Unicode compatibility and the ability to process input streams containing XML-encoded text, other types of data, and document metadata;
- embedding the RASP components within UIMA to improve scalability and interoperability (in collaboration with DigitalPebble Ltd) [R6]; and
- integrating the components with a machine learning classifier that is also distributed by iLexIR (http://ilexir.co.uk/media/langtech.pdf). The extended version of RASP is available under a commercial licence [see Section 5, C1].

Below we describe how the research underpins two commercial text-processing applications and services. We indicate the extent of the user population, and the economic benefits in terms of improvements to business processes and revenue from system development and licensing contracts.

- **Mobile-phone-based question-answering service**

  From 2004 to 2009, RE5ULT Ltd – under the trading names of 82ASK and then later Texperts – provided a UK-wide SMS question-answering service, employing human experts to answer questions submitted by the general public and charging £1 for each answer. In 2007, Texperts contracted iLexIR to develop an automated system to understand directory-enquiry-type questions, so they could be answered with minimal human intervention [C2].

  The system produced by iLexIR went live in February 2008. It takes an SMS text message as input and classifies it as being a directory enquiry or not; if it is, the system processes the text using RASP and (to the extent possible) extracts from the set of grammatical-relation analyses the type of enquiry (e.g. address or phone number), the type of entity, its full name, and its broad location. Back-end processing uses this information to query a directory database and generate an answer, which is checked by a human before being returned. The system ran until mid-2009, when Texperts was taken over by the US-based information services company kgb, and its operations subsumed into the parent company's existing infrastructure [C2].

  This kind of directory-enquiry automation had not been done previously in a commercial system. Reliably extracting such information from SMS text requires top-down information – embedded in the RASP system – about the grammar structures that the user is most likely intending to use, which are often obscured by misspelled, substituted or omitted words. The application would not have been possible without the underpinning research into the accurate parsing of text containing non-standard language and previously unseen words, and into non-deterministic language-processing strategies and grammatical-relation representations [C3].

  The Texperts service received around 100,000 messages a month from across the UK. Of the messages that were directory-enquiry-type questions, the automated system could extract useful information from 60%, and of these, over 90% were correctly interpreted [C3]. The time taken by a human to manually answer such questions was around 30 seconds, whereas the time taken for manual approval of an automatically answered question was only 5 seconds [C2]. Impacts were economic, through system development and licensing contracts to iLexIR worth over £50k [C4], as well as financial savings and service performance improvements for

Texperts.

- **The automatic grading of English-as-a-foreign-language examinations**

  Cambridge Assessment (CA) is one of the leading ESOL (English as a Second or Other Language) examining organisations in the world. More than two million people in over 130 countries take Cambridge ESOL exams each year. In April 2008, CA awarded iLexIR a contract to develop a system to automatically grade essay exams, working towards deploying the system as (1) an exam preparation aid that can instantly grade essays submitted online, and (2) an adjunct to human marking of exams, providing additional quality control and speeding up of assessment processes [C5].

  The automatic grading system is designed for CA's 'First Certificate in English', an upper-intermediate-level exam, one of whose components, 'Writing', requires candidates to compose a 200–400-word essay, which is graded on a scale of 1–40. The iLexIR grading system uses RASP to parse the essays, and then computes features relating to grammatical sequencing and structure (capturing correctness of grammar) and the complexity of sets of possible analyses (capturing information about grammatical sophistication), as well as features concerning word order. The features are passed to a machine-learning algorithm, which assigns a grade.

  Ablation tests demonstrate that the quality of automatic grading is much reduced if grammatical information is not used. Producing this grammatical information relies on several aspects of the underpinning research, in particular: accurate parsing of a wide range of formal and informal language usage, robustness to spelling mistakes, the efficient representation of the full space of possible analyses, and the recovery of fragmentary analyses from ungrammatical and non-standard language. An evaluation has shown that the system's grading is almost indistinguishable from that of experienced human examiners [C6].

  System development is leading up to imminent deployment as an exam preparation aid. The impact is economic, through system development, maintenance and licensing contracts to iLexIR worth a total of £450k over the REF impact period [C4].

As a further contribution to the economic impact benefiting iLexIR and establishing its viability, iLexIR has licensed the RASP system (or individual components of it) to three other technology-oriented companies and non-profit organisations during the REF impact period. These licenses were paid for in a combination of cash and equity with a notional total value of £90k, but with a current valuation much in excess of this [C4].

5. Sources to corroborate the impact

C1 http://ilexir.co.uk/licences-and-services/

C2 Communication from the former CTO, Texperts and VP Software Engineering, kgb.

C3 Confidential iLexIR report *82ask: DQ Performance and Error Analysis*; can be made available for audit purposes.

C4 Communications from the Company Secretary, iLexIR Ltd.

C5 Chief Executive, Cambridge English Language Assessment.

C6 Yannakoudakis, H., Briscoe, E. and Medlock, B. (2011) 'A new dataset and method for automatically grading ESOL texts', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 180–189, http://aclweb.org/anthology/P11-1019.