Q&A for computers

Description of proposed research and its context

If through some twist of fate you had to spend your life locked in the British Library, how much would you know about the world outside? *Data* is lack of uniformity within some context – and you would have no shortage of data, that's for sure. The books are as different as snowflakes. You might collect the various facts, stories, tables, and formulae they contain, and collate all of this into a truly gigantic encyclopedia. Perhaps as an expedient you would decide to bootstrap the process using an existing repository, like DBPedia, a crowdsourced database of facts extracted from Wikipedia. To test your comprehension of what you had studied, you would want to quiz yourself; and if you couldn't answer a particular question, you would have ample time to come back to it later. You could learn all manner of interesting things this way – but it would only get you so far. For example, it is quite clear that DBPedia and the British Library's 150m books will not tell you if the Riemann hypothesis is true, or not – because no one on Earth knows yet. If this catches your curiosity, you will need *information*. Information is what gives rise to form. The people who visit the library can teach you where the data that surrounds you comes from. Here, then, is what I propose. (1) Computer programs should be deployed to do what they are good at, which is churning away, processing and reprocessing texts, adding to and extending models. (2) People must be engaged with as well, and they can do what they're good at too, which is making meaning through social knowledge work – but with the addition of a reflective loop that asks *how* they go about it, in process terms. (3) These reflections should be integrated into the computational framework. (4) The computer system should undergo continuous testing in practical scenarios. As a result of following this strategy, the nominal person locked in the library – a stand-in for today's evolving computer systems – would not only understand a great deal about the world, but would also be able to interact with it in fruitful ways.

Background

There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. – Turing, 1951

Humans, for their part, are prolific conversationalists. There are 11m questions and 17m answers created by Stack Overflow users (technical computer programming Q&A); 532k questions and 766k answers on math.stackexchange.com (non-research mathematics); 66k questions and 106k answers on Mathoverflow (research-level mathematics). There are over 150 other sites in the Stack Exchange network, although not all of them are this popular. Mainstream social media is several orders of magnitude bigger (10b Facebook messages are sent *per day*). We are also quite good at communicating with computers in certain partial senses. There are 30m repositories on Github; and 5m articles on Wikipedia, from which 4.5m "things" and 580m "facts" are extracted into DBPedia. There are over a dozen popular natural language processing toolkits. This work is being augmented by other specialist efforts; e.g., ContentMine, a new project focused on extracting facts from scientific literature. The result is a growing machine-readable map of the world's knowledge as *content*. One example application is a recommender system that shows questions and answers from Stack Exchange to a programmer at work. Another largely distinct research subculture is devoted to mapping knowledge as *process* – the workings within the "social context" that facilitate knowledge production. Related research methods include social experiments in psychology, institutional analyses in economics, and various applications of network theory in sociology. My own work with the Peeragogy project uses design patterns – a heuristic method for capturing and expressing solutions to recurring problems that was originally pioneered in the field of architecture – to describe the way creative collaborations work. I am also involved with an effort to build a formal theory of the way informal collaboration works in mathematics, highlighting the domain-specific interaction of content and process. The next step in understanding the relationship between content and process could be made using simulation studies, targeting a "Stack Exchange for Computers"; a crucial step on the way to fulfilling Turing's vision of intelligent machines. As we learn more about the dynamics that relate content and process, we will be able to build *learning-aware* recommender systems that are able to deliver personalised tutoring; and, ultimately, *knowledge mining* systems that contribute directly to the growth of science.

National importance

Recognising higher-order patterns can help to cut costs while raising quality and controlling risk. This is relevant to everything from security to "smart services". The research proposed here will draw on a range of computer science methods to build a content-oriented model of technical domains, which will be directly useful for education and research. The most innovative feature of the proposal is that it will, in parallel, develop new methods to express process-oriented models of epistemic behaviour in computer programs, which can then be used to interact with, use, and expand this content. The Alan Turing Institute is an ideal context for this integrative work, which will in turn help the Institute fulfil its broader educational and outreach goals.

Academic impact

Working in an interdisciplinary manner will allow this project to collect and capitalise upon a wide range of low-hanging fruit, and promises further synergies. The reason this is important becomes clear with reference to mathematical artificial intelligence. Relative to the ambitions of computer scientists in the 1950s, understanding mathematical texts has proved slow in coming; robust NLP tools for mathematics are just becoming available now. Academic mathematics and computational "theorem proving" remain mostly disjoint. I have been involved in some preliminary steps to translate mathematical practice into computational terms. However, the current project proposes to tackle not just mathematics but other technical domains as well, which will enrich the inquiry with both technical and "common sense" features.

Research hypothesis and objectives

Referring to the definitions introduced earlier, the *data-oriented* aspects of the proposal focus on knowledge base construction, while the *information-oriented* aspects focus on modelling epistemic behaviours. The former will draw on relevant sources of open data (such as Wikipedia, Stack Exchange, the arXiv, Github), public domain materials and fair use of other relevant data sources. The latter will draw on direct observation, interviews, "instrumentation" of the social media accounts of researchers who agree to participate in the study, and software integration work as relevant. Findings will be added to a novel process-based model. The targeted advance is *machine understanding of knowledge artefacts and knowledge-producing processes*. This relates to several further hypotheses:

- We can progressively build text and domain understanding (from keywords found in the indices of books and triples found on the Semantic Web, to co-occurrence of themes in relevant corpora, to basic text understanding and generation).
- We can build computational models of social processes and research heuristics using a formal variant of the design pattern methodology (design patterns minimally express a *problem*, *solution*, and *rationale*; the precise formalism or combination of formalisms is to be determined and may vary with domain of application).
- We can use these formal design patterns to build high-level and low-level computational understanding of technical objects (including source code and commits in a variety of programming languages).
- The computer can itself contribute to these developments using code generation and a reflective model that conveys a degree of self-awareness, of "known unknowns" and other salient features.

Each of these hypotheses can be verified (or, in the event, refuted) in conjunction with practical work. A pilot project relevant to the first hypothesis would move from a collection of "named entities" gathered from the literature to a hierarchical outline of learning materials in technical subject areas. This would be carried out using the NNexus autolinking tool developed at PlanetMath to connect Wikipedia articles to relevant Q&A on Stack Exchange and to organise this material into learning pathways sorted by level of difficulty. Referring to the remaining hypotheses, existing agent-based systems for automatic programming and poetry generation can be adapted to carry out automatic authoring of design patterns and to develop high-level understanding of code. These tools would be deployed as bots offering criticism and code review on Wikipedia and Github, and deployed in a Q&A setting where they would "be able to converse with each other to sharpen their wits."

Programme and methodology

The primary research activities would be as follows, listed in order of "data-to-model-to-test"; as a matter of best practice, testing and integration would progress in rapid cycles.

A. Work with the BL collections and open online data to build a knowledge base. Employ off-the-shelf OCR software, standard NLP toolkits and data mining approaches. (≈20% time)

B. Anthropological research within ATI and partners, as a participant-observer in data science research, and as a participant-observer in online open source ecosystems. Use design patterns to build a collection of learning and research heuristics. (≈30% time)

C. Functional programming to encode the heuristics from **B** and connect with data from **A**. Tests of quality will progress through simulation studies, deployment to public-facing systems, and user studies. The software developed will be open source. (≈40% time)

This work would result in several major publications, as indicated in the following plan. Other publications would be developed along the way to these targets. Relevant conferences include AAAI, CSSSA, and SPLASH. (≈10% time will be reserved for professional development and dissemination activities.)

Paper 1 For a journal on knowledge discovery and knowledge bases. Summary of **A** at end of Year 1.

Bootstrapping knowledge-rich computing. Build a knowledge base by progressively moving from terms, to themes, to topics, to text understanding, drawing on open source knowledge resources and library materials. Individual technical terms can be taken as provisionally atomic, and relations between these atoms can be defined based on co-occurrence as well as grammatical structure. An empirical test of how well the computer understands textual content can be carried out by matching Stack Exchange questions to their answers, and by tagging duplicate questions.

Paper 2 For a journal on agent-based social simulation. Summary of **B** and **C** at end of Year 1.

The morpho-genetic basis of ideas in data science research. Use design patterns to capture and express research manoeuvres. Although design patterns have been employed extensively in the program *design* phase, this work will be novel in that the design patterns will be directly computationally meaningful. Empirical data and simulation studies will be employed; the twin aims are to use design patterns to build computer programs, and to use programs to write design patterns.

Paper 3 For a journal related to artificial intelligence. Combining **A**, **B**, **C** at the end of Year 2.

Machines conversing with each other to sharpen their wits. Use the programs and knowledge base developed in Year 1 to create a "Stack Exchange for Computers". Problems will posed and addressed by computational agents, and will be drawn from the real-world problems encountered by bots deployed on Wikipedia and Github as well as theoretical problems related to ongoing knowledge base construction and epistemic modelling. The system should demonstrate that computers can ask and answer questions about content and process. Feedback on bot behaviour will provide a contextual evaluation layer.

Paper 4 For a journal related to applications of artificial intelligence. Overall summary, end of Year 3.

The critically informed (automated) data scientist. Bridge the system from **Paper 3** to a platform where it can interact directly with human users, and where it can be critically evaluated by them. The aims are (i) to continue the system's education, both in terms of content and behaviour; and (ii) to investigate the degree to which interacting with the system supports learning and research outcomes for users. The system would be deployed to an in-house installation of the Open Source Q&A System (OSQA) or similar platform set up for discussion among ATI researchers. It would be rolled out to the standard Stack Exchange platform for a second study if its behaviour is deemed socially acceptable.

A possible continuation of this research into year 4 and year 5 of the fellowship can give an indication of the potential impact of this work. In this subsequent phase of research it would be appropriate to evaluate the computer's potential as a tutor, co-author, or author in its own right. That work would go hand-in-hand with a deployment that draws on the existing literature to model the growth and development of scientific and technical disciplines, starting with AI.