An AI-Inspired Intelligent Agent/Student Architecture to Combine Language Resources Research and Teaching

Bayan Abu Shawar, Eric Atwell

Information Technology and Computing Dept, Arab Open University, Amman, Jordan

and School of Computing, University of Leeds, UK

E-mail: b_shawar@aou.edu.jo, eric@comp.leeds.ac.uk

Abstract

This paper describes experimental use of the multi-agent architecture to integrate Natural Language and Information Systems research and teaching, by casting a group of students as intelligent agents to collect and analyse English language resources from around the world. Section 2 and section 3 describe the hybrid intelligent information systems experiments at the University of Leeds and the results generated, including several research papers accepted at international conferences, and a finalist entry in the British Computer Society Machine Intelligence contest. Our proposals for applying the multi-agent idea in other universities such as the Arab Open University are presented in section 4. The conclusion is presented in section 5: the success of hybrid intelligent information systems experiments in generating research papers within a limited time.

1. Introduction

For the UK Research Assessment Exercise (RAE) all UK academic researchers are required to produce 4 to 6 journal papers in about 6 years. This is crucial not just to build and maintain the research reputation of a university and academic department, but also for the promotion prospects of individual academics. In fact the academic promotion principles or rules are nearly the same in all universities worldwide. For example, in the Arab region, academic researchers need to publish 5-6 journal papers within not less than 5 years from their initial appointment, if they want a reasonable prospect of promotion.

Our hybrid of human natural language capability and intelligent information systems can produce 60 journal papers in 6 weeks - a BIG advance in Machine Intelligence AND Human Intelligence! MI+HI are integrated at 3 levels: use of AI architecture; intelligent use of MI tools; and student learning. Individual components are not particularly novel; the advance in MI comes through the novel combination. For the NLDB conference presentation, a volunteer from the audience can act as a student "intelligent agent", to demonstrate how AI architecture + student + MI can generate an outline journal paper.

Wikipedia explains that "In computer science, an intelligent agent (IA) is a software agent that exhibits some form of artificial intelligence that assists the user and will act on their behalf, in performing repetitive computer-related tasks. While the working of software agents used for operator assistance or data mining (sometimes referred to as bots) is often based on fixed pre-programmed rules, 'intelligent' here implies the ability to adapt and learn ... a multi-agent system (MAS) is a system composed of several agents, collectively capable of reaching goals that are difficult to achieve by an individual agent or monolithic system ... A multiple agent system (MAS) is a distributed parallel computer system built of many very simple components, each using a simple algorithm, and each communicating with other components. A paradigm of an ant colony or bee swarm is used many times..."

2. Multi-agent experiments

In a first experiment, Computational Modelling students were given the challenging yet clearly-constrained coursework task of developing and implementing a computational model for corpus based unsupervised machine learning of morphological analysis, for the PASCAL MorphoChallenge2005 research contest. Contestants were supplied with large corpus-derived wordlists for English, Finnish, and Turkish, one word per line. The challenge was to build a system which could read an input file, one word per line, and produce a corresponding output file with a space character inserted at every morpheme boundary. The students developed unsupervised learning algorithms for the English data set, which came with a test evaluation sample with morpheme boundaries marked; then went on to apply their unsupervised learning systems to the Finnish and Turkish word lists, this time without any knowledge of the target analyses. The outputs were sent back to the competition organisers for independent evaluation. The systems developed by the students ranged from minimalist to surprisingly successful; and we were able to combine these as components in a hybrid voting system, which performed better than any individual students' system [1]. This demonstrated that the Intelligent Agent architecture could be successfully applied to students. From a research perspective, a lone student cannot be expected to achieve great success, but the collective effort produced a range of systems, which when combined turned out to be as effective as other systems built by experts.

In a second experiment, Technologies for Knowledge Management and Computational Modelling students were given the data mining coursework task of harvesting and analysing a Data Warehouse from WWW, using web-as-corpus technology [2]. Each student/agent collected English language web pages from a specific national top level domain, and the analysis task involved comparing their national web-as-corpus with given Gold Standard samples from UK and US domains, to assess whether national WWW English terminology/ontology was closer to UK or US English. Results from 93 countries

worldwide were collated to give a collective response to the question: Which English dominates the World Wide Web, British or American? The task was cast as an exercise in applying the CRISP.DM methodology for computational modelling: the Cross Industry Standard Process for Data Mining projects. The CRISP-DM methodology specifies a series of phases or subtasks in a data mining project; it is a recipe to follow, allowing novices and non-experts to carry out data mining experiments successfully.

The students' success in carrying out the exercise is a testament to the practical value of the CRISP-DM methodology. The World Wide Web is divided into national domains, which makes it easy to collect a corpus of English language web pages from a specific country. Google has Advanced Search options to restrict results to a specified domain and language; WWW-BootCat uses Google to search for web pages, and allows users to specify these options. English is in effect a world-wide language on the WWW, in that a majority of web pages around the world are English (estimates differ, but it seems around 65% of web pages are in English). Most national domains, even where English is not a national language, include a large amount of English, showing that English is a truly international language. Our survey was not restricted to countries where English is a native language. We tried to include a wide variety of countries, and we succeeded in collecting 200,000-word samples from most national domains, together comprising approximately 20 million words of English from a world-wide spread of national domains.

The exceptions were either very small national domains with a very small population (e.g. South Georgia Island), or countries with legislation favouring a language other than English (e.g. Algeria has laws promoting publication in Arabic over ex-colonial French, and as a side effect these have also curtailed the use of English). Having collected their national English sub-corpus, each agent (student) had to decide whether it was closer to British or American English. Corpus Linguists on the CORPORA email discussion list forewarned us that the task would not be straightforward: many examples of supposedly American spellings are found in the British National Corpus, so we might have problems with the man-in-the-street assumption that these are two distinct varieties of English. As the students were substituting for AI intelligent agents (and as they were Computing rather than Linguistics students), they could not apply sophisticated linguistic knowledge to the problem.

Instead, each student/agent used simple computable measures to compare their national web-as-corpus with given Gold Standard samples from UK and US domains. The comparison methods included examining Log Likelihood profiles and averages comparing word frequencies in domain, UK and US corpora; counting occurrences of selected words known to have different UK/US spellings (eg color/colour); and counting occurrences of concepts realised by different UK/US words (eg fawcet/tap). Analysis was only at the lexical level: we had no means of comparing syntax or looking for characteristically UK v US grammar.

This exercise produced a detailed country by country analysis of the results from nearly a hundred student/agents, a large collection of national reports documenting the relative dominance of the two main varieties of English across the World Wide Web. However, although this exercise produced a large volume of results, it was still difficult to see patterns emerging.

As a follow-up experiment, Computational Modelling students were given the coursework task of explaining their computational modelling methods and results to an interdisciplinary journal readership, extending their results for their own national domain by comparisons with other students' findings for other countries in a geographical or political neighbourhood. The overall target was to showcase web-as-corpus data mining research methods to the wider language research community, by drafting research papers targeted at a range of new audiences, such as researchers in Middle Eastern Studies, Post Colonial Studies, Francophone Studies, English as a Foreign Language, English for Specific Purposes, and Language and Society.

3.   Results of Leeds experiments

As a follow-up experiment, Computational Modelling students were given the coursework task of explaining their computational modelling methods and results to an interdisciplinary journal readership, extending their results for their own national domain by comparisons with other students' findings for other countries in a geographical or political neighbourhood. The overall target was to showcase web-as-corpus data mining research methods to the wider language research community, by drafting research papers targeted at a range of new audiences, such as researchers in Middle Eastern Studies, Post Colonial Studies, Francophone Studies, English as a Foreign Language, English for Specific Purposes, and Language and Society. English to dominate the WWW: computing generally has been American led; and multinational companies with national branches might be expected to base their English-language pages on American originals. We were pleasantly surprised to find that UK English is holding its own on the WWW, and even preferred over US English in many domains and larger regions except for North and South America.

However, we also had an unforeseen finding: that often it was difficult to see any clear preference for British or American English, at least on the basis of the straightforward computational metrics available. Although intuitively there does seem to be a clear difference between the two varieties, in practice this actually affects only a very small proportion of words in web pages. The most noticeable difference between British and American English is in pronunciation, which of course is not apparent in web pages.

So far, we had collected web-as-corpus resources, leading to workshop and conference papers. Arguably the main deliverable of research is to publish journal papers; so we next embarked on another follow-up experiment, our most ambitious yet: this time, each student/agent had the task of re-using the bank of resources built so far by previous students/agents, to draft a large collection of research papers for submission to language-related journals, to demonstrate web corpus data mining research methods and results to a wider humanities research community. At the time of writing, the course has just finished, and the students have produced a bank of 60 draft research papers to polish before submission to 60 separate journals; the proofreading and polishing will be a

non-trivial, time-consuming task, but the rewards should be an impressive number of journal papers published in 2008/9.

In addition, we entered our system as a contestant in the British Computer Society's annual Machine Intelligence Prize contest. Our "Hybrid human-machine intelligence system to generate academic research papers" was a Finalist in the contest, at the BCS SGAI-2007 Conference at Cambridge University, December 2007 [4]. Unluckily we were beaten to First Prize by The Painting Fool, another system with a more photogenic demo; but the contest brought good publicity and interest from the conference participants.

4. Applying multi-agent architecture at Arab Open University

So far these techniques have only been tried at Leeds, but in principle this could be productive way of combining teaching and research more generally. For example, the Arab Open University (AOU) is a 6-year-old university which has 6 branches in 6 Arab countries, where it applies open learning techniques. The aim of the AOU is "to attract a large number of students who cannot attend traditional universities because of work, age, foundational reasons and other circumstances. The open terminology in this context means the freedom from many restrictions or constraints imposed by regular higher education institutions which include the time, space and content delivery methods [5].

The open learning strategy is considered as a hybrid between regular teaching and e-learning. Students have to attend face to face lectures once a week for two hours which is less then regular students. On the other hand, all module lecture notes, homework, etc. are found online for all students. In addition to that there are forums, and chat rooms to facilitate communication between students and their tutors, and students themselves.

In order to author our modules online and facilitating communication issues, AOU needs an e-learning platform which is" a software or a combination of software that sits on or is accessible from a network, which supports teaching and learningfor practitioners and learners" [6].

AOU has partnerships with the United Kingdom Open University (UKOU) and according to that at the beginning the AOU used the FirstClass system [7] as a computer mediated communication (CMC) to achieve good quality of interaction. The FirstClass tool provides emails, chat, newsgroups and conferences as a possible medium of communication. Nowadays AOU use Moodle [9] as an electronic platform, Moodle is an open source course management system (CMS) used by educational institutes, business, and even individual instructors to add web technology to their courses. To improve Moodle or even turn to using a new CMS, students who have the Telematics module which is a final year project could be recruited to work as a multi agents architecture. A proposal could be to divide these students into small groups that aim to collect information from the web about other CMS or e-authoring tools such as blackboard [8], Interact [10], CoMentor [11] and others. After that, a comparison study could be achieved to explore the best tool. Then students will find how to do some in-house development in Moodle to cope with our needs such as integrating the WAP and Mobile technology with CMS [12, 13], also utilizing some natural language tools to answer students' questions such as using chatbot. A chatbot is a conversational agent that interacts with users turn by turn using natural language. In this domain students may investigate the possibility of using a chatbot system in assisting teaching courses, learning new languages, and improving students' understanding for some modules by answering questions, conversation and sharing ideas [14-16].

In fact this study and results will not be only useful for AOU, but also for other public and private universities, which are moving toward using e-learning methodologies for some courses. At the same time many research papers could be generated about the platforms, comparisons studies, evaluation of usefulness, etc.

In another proposal, other students could be divided into three groups. The first and second group is responsible for collecting Arabic corpora from different web-pages using WWW-BootCat, each student in the group could collect a corpus from a specific Arabic country in a specific domain. Although all Arabic countries share the same written language form "Modern standard Arabic" in all its press, newspapers, however, none use the standard as its native spoken language [17]. So one group could collect corpora reflecting the standard Arabic language and other group collects different dialect corpora from chat rooms, forums, magazines. In combination with this, the third group could develop an Arabic morphological analyzer [18, 19]. Then apply this analyzer on both corpora the standard and the dialect ones to explore if the evaluation will be the same or not.

Unfortunately, till now there is a shortage of online Arabic language resources in comparison with other Language as English, this work could lead to develop big Arabic corpora at the end to enrich this field which could be used later by other researchers in develop online stemming lists, and POS taggers. In addition this will allow us to generate other journal and conference papers which compare between Arabic language used in different countries in variety domains and to filter the similarities between countries which belong to the same area like North Africa, the gulf, and middle east countries.

5. Conclusions

As well as achieving research goals, these experiments were novel and beneficial for student learning: they achieved the goal of research-led teaching and learning; student assessment was challenging and inspirational; and plagiarism was circumvented as each individual student task was novel and hence not easily capable. Student feedback was overwhelmingly positive: most relished the challenge of contributing to a real large-scale knowledge management data modelling task, and learning from hands-on experience of corpus linguistics research methods. The same ideas could be applied more widely in other universities; we are applying it at the Arab Open University to do hybrid intelligent information system research on virtual learning environments, e-learning platforms, and generating online Arabic language resources.

6. References

Atwell E; Roberts A. Combinatory Hybrid Elementary Analysis Of Text (CHEAT) *in: Kurimo, M, Creutz, M Lagus, K (editors)Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes.* 2006.

Atwell E.; Arshad J.; Lai C.; Nim L.; Rezapour Ashregi N.; Wang J.; Washtell J. Which English dominates the World Wide Web, British or American? *in: Proceedings of Corpus Linguistics 2007.*

Baroni M.; Kilgarriff A.; Pomikalek J.; Rychly P. WebBootCaT: instant domain.specific corpora to support human translators. In *Proceedings of EAMT 2006.* 11th Annual Conference of the European Association for Machine Translation. 2006.

British Computer Society. Machine Intelligence Prize, Summary of Finalists - 2007. [Online]: http://www.comp.leeds.ac.uk/chrisn/micomp/2007entries.html

Abu Shawar B. 2007. Utilizing AOU'VLE with other computerized systems. International Journal of computer Science and Security (IJCSS). Volume 1, issue 4. Pp 13-24.

http://www.elearningproviders.org/HTML/pages/link.asp

http://en.wikipedia.org/wiki/FirstClass

http://en.wikipedia.org/wiki/Blackboard_Learning_System

Nedeva, V. 2005. The possibilities of e-;earning, based on Moodle software platform. Trakia Journal of Sciences. Vol. 3, No. 7, Pp 12-19. www.interactlms.org

Gibbs, G.R. (1999). Learning how to learn using a virtual learning environment for philosophy. *Journal of Computer Assisted Learning.* Vol 15, pp 221-231.

Abu-Shawar B.; Al-Sadi J.; Sarie T. 2007. Integrating the Learning Management System with Mobile Technology, Accepted, WORLDCOMP'07, June 25-28, 2007, Las Vegas, USA, pp. 31-36.

Al-Sadi J.; Abu Shawar B. Using WAP Technology in E-learning, Accepted, WORLDCOMP'07, June 25-28, 2007, Las Vegas, USA, pp.41-46

Abu shawar B. and Atwell E. 2005. Using corpora in machin-learning chatbot systems. *International Journal of Corpus Linguistics* 10:4, pp. 489-516

Abu Shawar B; and Atwell E.. 2007 .Chatbots: Sind Sie wirklich nu"tzlich? (are they really useful?) *LDV-Forum Journal for Computational Linguistics and Language Technology*, 22(1) pp.29-49. 2007.

Abu Shawar B.; and Atwell E. 2004. Accessing an information system by chatting. In Meziane, F. and Metais, E. (eds.) *Proceedings of 9th International conference on the Application of Natural Language to Information Systems, NLDB 2004* Salford, UK. LNCS Lecture Notes in Computer Science, Springer, pp. 396-401

Atwell E.; Al-Sulaiti L.; Al-Osaimi S.; Abu Shawar B. 2004. A review of Arabic corpus analysis tools. In: Bel, B & Marlien, I (eds.) *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*. Vol. 2, pp. 229-234.

Emmam O., and Hassan H. 2003. Language model based Arabic word segmentation. In Proceedings of the 41st annual meeting of the association for computational linguistics, July 2003, pp. 399-406.

Habash N.; and Rambow O. 2006. MAGEAD: morphological analyzer and generator for the Arabic dialects. In *proceedings of Coling-ACL 2006, mail volume*. PP. 681-688.