|  |  |
| --- | --- |
| Title of project: | **A Web-as-Corpus approach to populating Wikiversity****for teaching about Islam and Muslims in language, linguistics and area studies** |
| Project Team: | Dr Eric Atwell, Prof James Dickins, Dr Majdi Sawalha |
| Institution: | University of Leeds |

**Final Report: December 2012**

Overview of project:

Undoubtedly a wealth of learning resources exists on the WWW, but scattered on individual websites, in a wide range of formats and structures. Wikiversity is an online open-source public repository for University-level teaching and learning materials, based on the Wikipedia architecture for “crowd-sourcing”: it relies on volunteers to collaborate by actively contributing their knowledge for the common good. Individual lecturers prepare online teaching materials to support their own teaching, but few know of Wikiversity, and few have time or inclination to take on the extra steps of formally organising and reformatting teaching materials for shared re-use. This project aimed to organise and semi-automate the harvesting of such scattered resources, by adapting Web-as-Corpus techniques from Corpus Linguistics. To demonstrate the feasibility of this methodology, we harvested from WWW sources a set of teaching materials for an example taught course module at Leeds University: ARAB5008M: Muslim Intellectual Encounter with Contemporary Thought.

Full description and outcomes:

Language and Linguistics research and teaching often involves a Corpus or empirical sample of the language being studied. The WWW is increasingly used as a corpus source, at three levels:

- A web search engine such as Google or Yahoo allows linguists to view the WWW as a corpus, by searching for and extracting empirical evidence of language items directly from the Web

- Scholars can “scout” the Web for suitable text-samples fitting their research needs, and download them to compile a specialised corpus, eg the Corpus of Contemporary Arabic (Al-Sulaiti and Atwell 2006)

- A “robot” can be used to automate the process of harvesting a corpus fitting specified criteria; for example BootCat is given a list of specific “seed terms” and uses these to harvest WWW documents matching some or all of these seed terms, eg (Atwell et al 2009, Atwell et al 2011)

We investigated and adapted these three complementary approaches, to harvest University-level teaching and learning resources for teaching and learning about Islamic Studies. This involved:

- Investigation into search strategies and terms to identify candidate teaching materials appropriate to Islamic Studies university courses

- “scouting” for resources: applying these search strategies and terms to find and download a “candidate corpus” of web-based resources which appear potentially useful for Islamic Studies teaching and learning

- Investigation of the possible use of BootCat, the web-as-corpus robot, to automatically harvest potential Islamic Studies teaching materials; for this, we tried to identify seed-terms and parameter constraints which enable BootCat to target potentially suitable websites.

In parallel to this three-pronged approach to “harvesting”, we developed an evaluation procedure, to filter out unsuitable candidate teaching resources. Potential resources were reviewed by academic teaching staff in Arabic and Middle Eastern Studies, who used their expertise and judgement to evaluate (i) relevance to their own teaching, and (ii) potential usefulness to the wider community of teachers and learners about Islam and Muslims.

This could be a programme for a large-scale research and development project; but given the limited budget of £3000, we were restricted to a pilot feasibility study. We hired a temporary Intern: Majdi Sawalha, who is a talented Computing and Artificial Intelligence student employed for a short-term pilot study to adapt the three web-as-corpus techniques to our task of harvesting candidate teaching materials. This technical development and corpus harvesting was supervised and guided by Atwell, leader of the Arabic Computing research group and experienced with web-as-corpus techniques for Corpus Linguistics research, e.g. (Al-Sulaiti and Atwell 2006, Atwell et al 2009, Atwell et al 2011). Our pilot study focussed on a restricted target “application”: teaching and learning about Islam and Muslims in the School of Modern Languages and Cultures at Leeds University. The evaluation of candidate resources focussed on potential for reuse in specific taught modules in this School. Dickins is Professor and Head of Department of Arabic and Middle Eastern Studies and Director of the Language, Linguistics and Translation research group within the School of Modern languages and Cultures; Dickins and his colleagues evaluated candidate teaching materials for potential reuse in the department’s teaching. We identified an example taught course module at Leeds University, to use as a case study: ARAB5008M: Muslim Intellectual Encounter with Contemporary Thought. Resources judged potentially useful were followed up, and collated into a teaching resource pack.

The main outcomes of the project are:

* A collection of online reusable teaching materials for Islamic Studies, specifically relevant for the taught course module at Leeds University ARAB5008M: Muslim Intellectual Encounter with Contemporary Thought – see Appendix A.
* A methodology for harvesting further resources, reusable for other courses and universities.
* An empirical evaluation of this methodology and its applicability to teaching and learning about Islam and Muslims; we identified high-quality reusable teaching resources for the case-study taught module.
* A report on the pilot study results and lessons learnt, to be submitted to an appropriate journal (to be decided) – see Appendix B.
* Inspiration to help us plan a follow-on research grant proposal to the Engineering and Physical Science Research Council (EPSRC) for a 350K grant for a project on “Natural Language Processing Working Together With Arabic and Islamic Studies; this grant has been awarded to Leeds University, for a 2-year project to run from February 2013 to January 2015.

Benefit of the project to the wider Islamic Studies higher education community:

The outcomes listed above are of themselves self-evidently beneficial:

* The collection of online reusable teaching materials for Islamic Studies can be reused widely, not just at Leeds University
* The methodology and tools for harvesting further resources can also be reused more widely, again not just at Leeds University
* The evaluation of the methodology will enable us (and others) to decide whether to proceed further, or else perhaps to abandon the idea as impractical.
* Whatever the outcome, publication of the pilot study in journal(s) is a benefit.

Other benefits include:

* Enhanced awareness of online resources and the benefits of reusability
* Collaboration between Language and Computing researchers and scholars, to mutual enrichment
* The new perspective on Islamic Studies via online reused resources should add variety, novelty and hopefully inspiration to the student learning experience.

Innovation:

Applying Web-as-corpus harvesting techniques and technology to Islamic Studies is clearly novel. Some other projects funded by HEA enable individuals or groups to develop teaching resources for their own use; reusability of resources is central to Software Engineering, but less commonplace in teaching about Islam and Muslims.

The “gap” we aimed to fill is not a specific subtopic in Islamic Studies, but a gap in reuse and sharing of learning resources generically.

The project is also innovative in terms of Computing research: to our knowledge, Web-as-Corpus techniques have not been previously adapted to harvesting teaching materials. This could open up a new area of applied research in Computing and Artificial Intelligence.

Dissemination

Wikiversity is an established architecture for sharing and dissemination of teaching resources, which may persuade Language, Linguistics and Area Studies scholars that the resources are ready and easy to re-use. The novelty of our approach to “harvesting the web for Islamic Studies” should also help grab attention, and inspire learners and teachers to try our results. The experimental case study is publishable as computing research, and this should help publicise the Islamic Studies learning resources as a by-product.

Monitoring and Evaluation

In our post-project publications, we can report on the number and range of teaching resources harvested, and the range of taught courses serviced by these harvested resources. This evaluates effectiveness in terms of usefulness for teaching and learning about Islam and Muslims. We can also apply metrics from Machine Learning and Information Retrieval research (precision, recall, etc) to the software tools, such as our adapted version of BootCat. The most important outcome will hopefully be reuse of our contributions to web-based teaching resources, though this can only be judged in the longer term. Another outcome has been to seek funding for a larger-scale follow-up research and development project: we applied to the Engineering and Physical Science Research Council (EPSRC) for a 350K grant for a project on “Natural Language Processing Working Together With Arabic and Islamic Studies; this grant has been awarded to Leeds University, for a 2-year project to run from February 2013 to January 2015.

Costs and Value for money

Given that only £3000 was on offer, the best use of this limited sum was to hire a student Intern: Majdi Sawalha, who is a talented Computing and Artificial Intelligence student employed to develop the methodology and use it to harvest potential resources. Ideally the supervision and evaluation contributions of Atwell and Dickins ought to have been funded as staff buy-out time; but £3000 was barely sufficient to fund the student intern, so Proposers’ time had to be contributed by Leeds University. The student Intern was also provided with use of desk, PC and research computing infrastructure in our AI Lab, to work alongside externally-funded postgrads and postdocs. So, this was excellent value for money: research on the cheap! Note that the follow-on EPSRC-funded project will cost 350K, and this will cover 80% of Full Economic Costing of the project, including 5-10% of Investigator time.

References

L Al-Sulaiti and E Atwell. 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171

E Atwell et al. 2009. Arabic and Arab English in the Arab World. Proc CL2009 International Conference on Corpus Linguistics, Liverpool University, UK

E Atwell et al. 2011. An Artificial Intelligence approach to Arabic and Islamic content on the internet. Proc NITS'2011 National Information Technology Symposium, King Saud University, Saudi Arabia.

**APPENDIX A: A collection of online reusable teaching materials for Islamic Studies, specifically relevant for the taught course module at Leeds University ARAB5008M: Muslim Intellectual Encounter with Contemporary Thought**

**APPENDIX B: Draft journal paper reporting on the project**