

Combining Teaching and Research in Text-Mining from Social and Cultural Data.

Claire Brierley, Bolton University and University of Leeds
Eric Atwell, University of Leeds

We propose a potentially rich resource for e-Social Science text mining research. Many Computing students undertake “mini-projects” in text-mining; if they are appropriately organised and supplied with Social Science text-data to mine, the results can be significant, particularly if outputs can be coordinated and combined. We describe how to use a multi-agent architecture from Artificial Intelligence to integrate research and teaching in text-mining from social and cultural data, by applying Computing students as “intelligent agents” to collect and mine social data.

Our first attempt to coordinate a class of students to act as a set of “intelligent agents” was for an unsupervised Machine Learning linguistic analysis challenge: to mine data-sets of English, Finnish and Turkish texts and learn rules for morphological analysis of words in each language. The individual results of “agents” (students) were combined into a multi-agent solution which achieved higher accuracy than any one individual rule-set (Atwell and Roberts 2006).

The next challenge was to investigate socio-cultural influences on the World Wide Web; specifically, focussing on whether American or British influences dominate the Web. Each student collected and analysed English-language Web-pages from a national Top Level Domain, eg .AR = Argentina, using the WEKA Data-Mining toolkit to find empirical evidence of British and/or American influence. The results of the overall exercise include: an International Web-Corpus of English: 200K-word samples from c100 national domains; c100 research reports, showing whether UK or US influences predominate in each national domain; summaries for larger “regions”, eg Middle East, Francophone region, South-East Asia; 6 student papers accepted for the International Conference on Corpus Linguistics 2007; and a Finalist in the AI 2007 British Computer Society Machine Intelligence Contest (Atwell et al 2007, Atwell and Abu Shawar 2008).

These tasks involved students explicitly collecting web-sourced data before going on to investigate it via data-mining. Computing students also generate personal data (eg through project logs). Brierley is student projects coordinator at Bolton University, and coordinates the information infrastructure for student project management; we plan to investigate this as a potential data-source for text-mining analysis of social interaction at the micro- rather than macro-level. Ideally students will not only be data-sources, but also “intelligent agents” for processing the data, applying text-mining tools; of course privacy and security will be vital.

Our overall aim is to combine teaching and research to mutual benefit. Hybrid student-machine “intelligent agents” can be used to generate e-Social Science Text-Mining research within limited resources and time.

References

Atwell E; Roberts A. 2006. 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.

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

Atwell E, Abu Shawar B. 2008. An AI-Inspired Intelligent Agent/Student Architecture to Combine Language Resources Research and Teaching. In: Proceedings of LREC Language Resources and Evaluation Conference.