Combinatory Hybrid Elementary Analysis of Text

Eric Atwell[1]

Abstract

We propose the CHEAT approach to any corpus analysis or annotation challenge: Combinatory Hybrid Elementary Analysis of Text. The idea is: sit back and get others to do the work, then simply learn from their results.

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."

This paper describes experimental use of the multi-agent architecture to combine research and teaching in corpus linguistics, by casting students as intelligent agents.

In a first experiment, Cognitive Systems and Multidisciplinary Informatics MSc students in my Computational Modelling class were given the challenging yet clearly-constrained coursework task of developing and implementing a computational model for corpus-based morphological analysis, for the PASCAL MorphoChallenge2005 research contest. The systems developed 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 (Atwell and Roberts 2006).

The first step in our CHEAT approach was to acquire results from a number of other candidate systems which attempt the task. We then developed a simple CHEAT program, to read in the output files of each of the other systems, and then line-by-line select the "majority vote" analysis - the analysis which most systems have gone for. If there is a tie, it takes the result produced by the system with the highest F-measure; if the other systems' output files are ordered best-first, then this is achieved by simply taking the first of the tied results. To demonstrate our approach, we entered our CHEAT system in the MorphoChallenge contest for unsupervised-learning morphological analysis systems, alongside the individual entries from students. All entrants had to attempt morphological analyses of English, Finnish and Turkish corpus-derived wordlists.

We were able to argue that CHEAT is more than just ordinary unsupervised Machine Learning: the CHEAT approach involves super-sized unsupervised learning!

[1] University of Leeds
e-mail: eric@comp.leeds.ac.uk

CHEAT combines not just one or two but three different layers of unsupervised learning: (1) Unsupervised learning by autonomous agents: we cast our Computing students as agents, requiring them to write a program for coursework; (2) Unsupervised learning by the set of student programs; and (3) Unsupervised learning by our own CHEAT Python program, cheat.py, from the output of the set of student programs.

For all three languages (English, Turkish, Finnish), our CHEAT system scored a higher F-measure than any of the contributing systems developed by our Leeds students. It also achieved better Precision and Recall scores, with a couple of exceptions. The MorphoChallenge workshop organizers went on to see if CHEAT would also work with the best systems overall from all entrants to the contest. They took our CHEAT.py program and used it to combine results from the top 5 systems; they found that the results were again better than the output of any one individual system: if we had been allowed to "go back in time" and enter this hybrid system, we would have easily won the contest!

Clearly, Combinatory Hybrid Elementary Analysis of Text is a valid approach to Unsupervised Learning of morphological analysis, and it should be readily adaptable to other corpus analysis tasks, so long as we can get others to cooperate and give us their results. However, a limitation of our basic CHEAT.py program is that it requires all results to be parallel, aligned output files, with straightforward unlabbeled analysis such as insertion of space to mark morpheme boundary. This was a problem for the second MorphoChallenge contest, where entries had to also add morpheme labels. Each entrant could come up with their own labels, making entries mutually incompatible: each "vote" could be different even if morpheme-boundaries were the same.

In a second experiment, we changed the "agent architecture": instead of getting each "agent" to carry out the same analysis, and then combine the results by a voting system, we tried a "massively parallel architecture": divide the processing into individual subprocesses, get each agent to process their own separate data, and then combine the results. Final-year Undergraduates and MSc students in two of my classes, Technologies for Knowledge Management and Computational Modelling, were given the data-mining coursework task of harvesting and analysing a Data Warehouse from WWW, using web-as-corpus technology (Baroni et al 2006). 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 gathered, to give an overview answer to the question: Which English dominates the World Wide Web, British or American? (Atwell et al 2007).

Our third experiment is the most ambitious to date: 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 is to publish corpus linguistics research papers in a range of journals, introducing web-as-corpus methods and results to new audiences, such as journals of Middle Eastern Studies, Post Colonial Studies, Francophone Studies, English as a Foreign Language, English for Specific Purposes, and Language and Society.

As well as achieving research goals, these experiments were novel and beneficial for student learning: they achieved the Leeds University 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 copyable. 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.

References

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

Atwell, Eric, Junaid Arshad, Chien-Ming Lai, Lan Nim, Noushin Rezapour Asheghi, Josiah Wang, and Justin Washtell. Which English dominates the World Wide Web, British or American? Submitted to Corpus Linguistics 2007.

Baroni, Marco; Kilgarriff, Adam; Pomikalek, Jan; Rychly, Pavel. 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.