Tidying up the Basement: A Tale of Large-Scale Parsing on National eInfrastructure

Stephan Oepen
Universitetet i Oslo
oe@ifi.uio.no

ABSTRACT
Until about six years ago, our research group used non-trivial amounts of project funds and researcher time on maintaining a dedicated server farm in the basement of our department. Rack space and cooling (just as much as funds and time) were in short supply, and we never quite got around to implementing automated load balancing across compute nodes, tuning the Linux kernel and filesystem for optimum performance, or connecting to the uninterruptible power supply. When pointed to the Norwegian National High-Performance Computing Initiative, we were intially doubtful that Natural Language Processing should be among their target user groups. Also, we were a tad hesitant to give up control of our own equipment and of course worried we would miss what we thought were our fancy toys. Today, any member of the group can access thousands of cpus simultaneously, we have about five terabytes of project data on-line, and our research has scaled to dataset sizes and turn-around times that would be just inconceivable on group-local hardware—at no charge to our project funds and no administrator responsibilities. For example, 'deep' semantic parsing of the about 900 million words of the English Wikipedia we can typically complete in less than one day (while expending what would be about eight sequential years of computation). Or, when searching for the best-performing features and hyper-parameters in a machine learning problem, we can explore a large 'grid' of possible configurations in parallel, without much need for a staged, partly manual, 'coarse-to-fine' search strategy. Access to the very large-scale Norwegian National eInfrastructure and its high-quality technical support have enabled a comparatively computation-heavy research profile of our group and has thus contributed to its international competitiveness. In this presentation, I will review some of our experiences in establishing a dialogue with the HPC crowd and propose *HPC for the Masses* as a candidate vision in the on-going development trend towards more and more large-scale computational sciences.

KEYWORDS: Parsing Wikipedia, HPC for the Masses, National eInfrastructure.