Stan Matwin: “Experience with a Deployed Text Mining System”

Abstract:

One of the basic tools of the Evidence-based Medicine is the so called Systematic Review. The first step in the development of Systematic Reviews is manual classification of a large corpus of abstracts of medical articles into two classes: those relevant, and those not relevant to the topic of the study. We have applied Machine Learning techniques to this task, in a joint project with a company specializing in Systematic Reviews. Our learning solution is incorporated as an option into its product. We will discuss how we progressed through the project, focusing on significant challenges that we have encountered in what seemed initially to be a straightforward text classification exercise. These challenges are due to the characteristics of the corpus, the inadequacy of standard text classification techniques, the need for non-standard text representation, the need for non-standard evaluation measures, specific performance requirements, etc. Resulting “lessons leaned” can be somewhat generalized to other applied and deployed AI projects.