Dataset Profiling: Exploring Data Sensitivity in Corpus Linguistics

Presenter: Anne de Roeck

In Corpus linguistics and Information Retrieval, it is an inescapable fact of life that, given a task, the performance of a technique will depend on the properties of the language data on which it is run. Unfortunately, whilst data and language sensitivity is often acknowledged, the systematic exploration of the relationship between data, task type and technique performance is absent. This leads to a number of methodological and practical problems - for instance because experimental results are not replicable on different data. The current emphasis on corpus driven experimental work suggests that this is an area in need of investigation.

This talk starts by demonstrating the impact of different kinds of language data and showing that corpora do differ significantly in ways which are known the affect experimental results. Different languages have different characteristics which affect the success with which statistical techniques can be applied. One way of capturing salient differences between corpora or datasets is via the development of measures that bring out the bias a collection may have with respect to a particular technique. Data profiling is introduced a an approach in which the distinctive characteristics of a collection can be captured, and some measures are proposed that may be useful in building profiles.