Computational Linguistics could Serve History of Science

John Nerbonne
Computational Linguistics, University of Groningen

While it is true that Computational Linguistics (CL) is largely focused on the refinement of technique, as the general description of this meeting noted, there are a number of strands of current CL work that could serve the study of the history of science. This would be fitting, as one of the greatest advances in textual information retrieval has been the incorporation of scientometric techniques.

CL began with ambition to study language from a computational perspective, suggesting that the goal was to contribute to Linguistics, but CL contributions to Linguistics have not been well received, leaving the field increasingly focused on CL applications, e.g. in text search (information retrieval), computer-assisted language learning, automatic question answering, machine-aided translation, automatic summarization, and many more.

Terminology extraction seeks to identify not only the significant vocabulary in a scientific field, but also the hierarchical sub-sort relation among the elements denoted by the terms (what is now termed the "ontology" of the field). It is not a solved problem, but progress is measurable and real, yielding results such as "streptococcal pneumonia is a kind of bacterial pneumonia is ... a kind of cardio-pulmonary disease ..." Similar ontologies are inferred for treatments and symptoms. Bootstrapping on terminology extraction, relation extraction infers relations between elements of ontologies such as the relation between a malady and a symptom or that between a treatment and a side effect. The success of such studies are tested in the degree to which (true) scientific knowledge is extracted from text.

While there appears to have been little work linking history of science to CL research of the sort sketched above (or of other sorts we can indicate briefly), perhaps there is a bit which we can note, and there is room for interesting experimentation.