Automatic fact extraction for simulation purposes

Presentation of a software toolbox that automatically identifies and extracts qualitative and numerical values from scientific texts.

H. Busch, L. Barrio-Alvers, C. Zinecker, M. C. Hirsch;

interActive-Systems GmbH, Scheppe Gewissegasse 28, D – 35039 Marburg

Correspondence to: Heiner.Busch@interActive-Systems.de

It is commonplace that the number of publications in the biomedical literature is increasing exponentially. For scientists it is thus difficult to stay up to date in their field of interest let alone to stay informed interdisciplinary. In the context of numerical simulations of biological processes interdisciplinarity is a prerequisite, because one needs to integrate data from basic as well as from clinical research.

On the other hand we experience an equal acceleration in the amount of available computing power, as well as an increasing number of papers published in electronic form in addition to the printed version. Albeit most of the published data is still unstructured prosa text and at best poorly annotated with metadata. This has led to many national and international initiatives that strive at using computers to automatically analyze scientific texts in order to extract, annotate, structure, reorganize and present those facts which are relevant for the context in which a researcher has an information need. This is an overwhelming endeavor in the light of the number of biomedical disciplines and their diverse and domain specific vocabularies. In order to reduce it to a feasible and provable workload and in order to serve the European network of excellence “BioSIM”, we focused on the domain of numerical simulations of biological processes.

Our approach is strictly focused on the identification and extraction of qualitative, quantitative and numerical parameters and their respective values. These parameters need not all be known in advance, but the software will identify and present new candidates of such parameters, which can then be fed back into a learning loop for future fact extraction.

In our poster we will present our computational workflow, its components and first results.

While the identification of relevant data in unstructured text is well under way and results are promising, it is still difficult to computationally extract the data from tables, figure legends or even figures.

Bundesministerium
für Bildung
und Forschung

This project is funded by: BMBF and EU FP6