



Dietrich Rebholz-Schuhmann

PhD in immunology, University of Düsseldorf, 1989. Senior scientist at gsf, Munich, 1995. Director Healthcare IT, LION Bioscience AG, Heidelberg, 1998. At EMBL-EBI since 2003.

Literature Research

DESCRIPTION OF RESEARCH

Text mining comprises the fast retrieval of relevant documents from the whole body of the scientific literature and the extraction of facts from these texts. Text-mining solutions are becoming mature enough to be automatically integrated into workflows for research and into services for the general public, for example delivery of annotated full text documents as part of UK Pubmed Central (UKPMC).

Research in the Rebholz group focuses on extracting facts from the literature. Our goal is to connect literature content automatically to other biomedical data resources and to evaluate the results. On-going research targets the recognition of biomedical terms (genes, proteins, gene ontology labels) and the identification of relationships between them. Our work is split into three tightly coupled parts: named entity recognition and its quality control (e.g. UKPMC project); knowledge discovery (e.g. identification of gene–disease associations); and further development of the IT infrastructure for information extraction and fact delivery.

SUMMARY OF PROGRESS

- Further developed different solutions to normalise the representation of concepts in the literature: LexEBI, IeXML, Whatizit and CALBC (Collaborative annotation of a large-scale corpus) and all solutions contribute to the annotation and indexing of the full text scientific literature as part of the UKPMC and the SESL project;
- Finalised an ontological framework for phenotypic descriptions that offers new possibilities to represent disease phenotypes and and compare them (Hoehndorf et al., 2010);
- Developed and evaluated new solutions for the characterisation of protein splice variants from the literature and the extraction of gene regulatory events;
- Processed full text documents from major publishers – as part of the SESL project – to integrate the extracted evidences with bioinformatics data resources (e.g. UniProt, ArrayExpress) and to deliver the assertions from the SPARQL endpoint to the project partners in the pharmaceutical industry;
- Launched the Journal of Biomedical Semantics (JBMS).

MAJOR ACHIEVEMENTS

Named Entity Recognition

Standardisation of the scientific literature: UKPMC, LexEBI, and CALBC

Delphine Bas, Adam Bernard, Abhishek Dixit, Senay Kafkas, Jee-Hyub Kim, Vivian Lee, Ian Lewin, Chen Li, Maria Liakata, Menaka Naraysamy, Piotr Pezik, Rohit Rexa, Shyamasri Saha, Dolf Trieschnigg, Ying Yan, Antonio Jimeno Yepes

Our research focuses on identifying named entities (e.g. genes, proteins, diseases) from the literature and linking them to an entry in a reference database. Several solutions have been provided: LexEBI (a terminological resource), IeXML (an annotation framework for documents), Whatizit (an information-extraction infrastructure) and CALBC (an evaluation infrastructure). We are generating LexEBI in order to provide full coverage of domain knowledge in molecular biology for gene and protein names, chemical entities, diseases, species and ontological terms. During the reporting period several bioinformatics resources were incorporated into LexEBI (e.g. the BioThesaurus), which interlinks terms across all resources according to their similarity. LexEBI supports large-scale information extraction in the biomedical domain and has been integrated with IeXML and the EBI's information-extraction infrastructure for indexing of the full body of scientific literature (UKPMC project).

By harmonising outputs from automatic text mining solutions, CALBC project partners produced SSC-I, a large-scale, annotated biomedical corpus containing annotations from four semantic groups (chemical entities and drugs; genes and proteins; diseases and disorders; and species). SSC-I has been used for the First CALBC Challenge, wherein participants were asked to annotate the corpus with their annotation solutions (Rebholz-Schuhmann, Jimeno Yepes et al., 2010). As of July 2010, SSC-I delivers

1 121 705 annotations for 100 000 Medline abstracts. The annotations are sufficiently homogeneous to be reproduced with a trained classifier (F-measure of 85%).

KNOWLEDGE DISCOVERY

Novel solutions for discourse analysis

Maria Liakata, Shyamasri Saha

We have created an annotation scheme (CoreSCs) that distinguishes the following core categories within the discourse of a scientific publication: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion. We have asked chemistry experts to use this scheme to annotate a corpus of 265 full papers in physical chemistry and biochemistry. We are training and testing machine-learning algorithms for the automatic classification of sentences in papers according to this annotation scheme. An SVM classifier featuring a linear kernel and a Conditional Random Fields (CRF) classifier achieve 48% and 50% accuracy, respectively. We achieved the best accuracy for Experiment and Background.

Development and use of phenotype ontological resources

Robert Hoehndorf, Anika Oellrich

The use of bioinformatics data in clinical environments requires the consistent and complete representation of phenotypes. For example, in DECIPHER (which uses Ensembl resources) a patient or syndrome phenotype is expressed with a specific subset of the London Dysmorphology Database (LDDDB) terminology. We formalised a framework for phenotypic descriptions that makes their semantics explicit, thus providing the means to integrate phenotypic descriptions with ontologies of other domains and offering a new capability to represent disease phenotypes and perform powerful queries (Hoehndorf et al., 2010).

Improving the extraction of complex regulatory events from scientific text using ontology-based inference

Jung-Jae Kim

The extraction of complex events from biomedical text requires in-depth semantic analysis. We developed a system that deduces implicit events from explicitly expressed events using inference rules that encode domain knowledge. We evaluated the system with the inference module on different tasks and found that the inference based on domain knowledge plays a significant role in extracting complex events from text. This approach has great potential to recognise the complex concepts of biomedical ontologies (e.g. Gene Ontology) in the literature.

IT infrastructure development for information extraction

The SESL Triple Store: retrieval over large literature content

Samuel Croset, Christoph Grabmüller, Silvestras Kavaliauskas, Chen Li, Darius Sulskus

As part of the SESL project, the Pistoia Alliance (in collaboration with Nature Publishing Group, Elsevier, Oxford University Press and the Royal Society of Chemistry) produced an RDF Triple Store representation to simultaneously query the publishers' content and bioinformatics data resources using the RDF query language (SPARQL). The Triple Store delivers gene–disease associations from the scientific literature to the users through several interfaces: SPARQL queries, SOAP web services and a graphical user interface. The Triple Store contains about 14.5 million triples from the scientific literature that have been aligned with content from the Gene Expression Atlas (182 840 triples) and UniProtKb (12 552 239 triples for human). The RDF Triple Store enables simultaneous querying of the scientific literature and bioinformatics resources for evidence of gene–disease links.

SELECTED REFERENCES

Hoehndorf, R., Oellrich, A. and Rebholz-Schuhmann, D. (2010) Interoperability between phenotype and anatomy ontologies. *Bioinformatics* (in press). Published online 22 October; DOI: 10.1093/bioinformatics/btq578.

Rebholz-Schuhmann, D., E., et al. (2010) CALBC Silver Standard Corpus. *J. Bioinform. Comput. Biol.* 8, 163-179.

Rebholz-Schuhmann, D. and Nenadic, G. (2010) Biomedical Semantics: the Hub for Biomedical Research 2.0. *J. Biomed. Semantics* 1, 1.

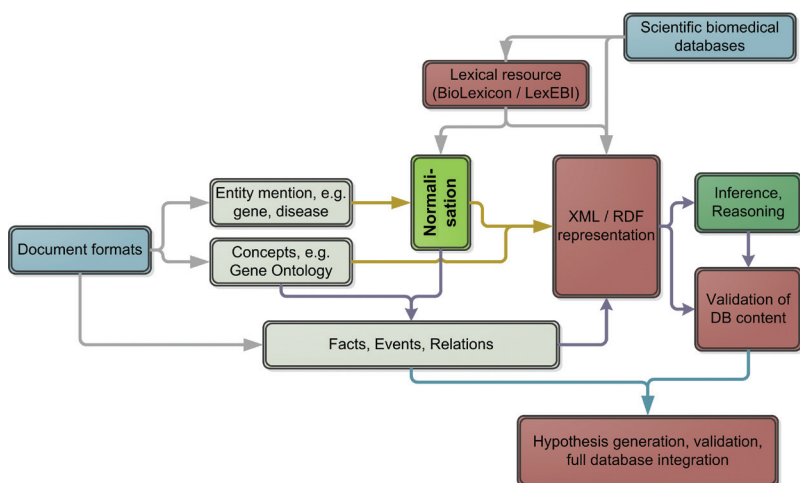


Figure: Literature analysis analyses scientific documents, identifies entities, concepts and facts (grey boxes) and normalises the entities to database entries with the support of a lexical resource (BioLexicon / LexEBI). RDF representations of the facts in combination of ontological resources supports inference and reasoning across the data content.