Ontologies and Text Mining for Life Sciences
Current Status and Future Perspectives

Dagstuhl, 25-28 March 2008

Executive Summary

Keywords: Text Mining, natural language processing, ontologies, ontology design, machine learning, bioinformatics, medical informatics, knowledge management

1 Introduction

Researchers in Text Mining and researchers active in developing ontological resources provide solutions to preserve semantic information properly, i.e. in ontologies and/or fact databases. Researchers from both fields tend to work independently from each other, but there is a shared interest to profit from ongoing research in the complementary domain. The relatedness of both domains has led to the idea to organize a workshop that brings together members of both research domains.

2 The gap between Text Mining and ontologies

Life Science researchers deliver their findings in scientific publications. These documents are nowadays distributed electronically and increasingly processed by automatic means to also incorporate those findings and the data into structured, scientific databases. Methods for this purpose are generally subsumed under the term “Text Mining”, encompassing techniques belonging to the fields of machine learning, information retrieval and natural language processing. Text Mining-based solutions have, for instance, been developed for the identification of protein-protein interactions, of gene regulatory events, for the functional annotation of proteins, for the identification and prioritization of disease-related genes, and for the analysis of results from high-throughput experiments.

Text Mining for the Life Sciences has received considerable interest over the last years and is now an established area for conferences and workshops (e.g., ISMB, KDD, ECCB, Coling, ACL, PSB) and has lead to international large-scale challenge events (KDD-Cup, Genomics track at TREC, BioCreative2&2, BioNLP). The cause for this interest is the ever increasing amount of publications imposing an unbearable work burden on the individual researcher and the promising advances in natural language processing and machine learning that form the solution to the problem, if they are integrated into biomedical applications.

Text Mining has to cope with a large semantic gap between the raw textual data and the representation of meaningful results in databases, e.g., normalization of events in the text to conceptual representations of events according to “textbook” knowledge. It is hoped that ontologies fill this gap delivering a structured representation of biomedical knowledge. Although large and increasingly comprehensive biological ontologies

Dagstuhl Seminar Proceedings 08131
Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives
http://drops.dagstuhl.de/opus/volltexte/2008/1523

are now available for many relevant topics (e.g. Gene Ontology, Sequence Ontology, Phenotype Ontologies etc.), it has not yet been proven what type of resources are ideally suited for Text Mining solutions.

Investigating on the aims of research in Text Mining and in ontological design, we find that ontologies are not designed to support Text Mining but rather to improve the annotation of database content. Although, Text Mining solutions intend to fill databases with content, it is not the case that Text Mining solution find ontological concepts easily in the literature, and, even more, ontological resources are not designed to support Text Mining solutions in the sense that the ontological terms fit to the demands of a natural language processing system. However, the Text Mining community exploits ontological resources to link generated evidence from the literature to the ontological concepts. Furthermore, the ontologies are not only a tool, but also a target for Text Mining research. Plenty of methods have been devised that automatically or semi-automatically construct ontologies or enrich existing ontologies by extracting terms and relationships from biomedical text collections.

These areas are researched by a community of researchers working in a highly interdisciplinary way in the domains of biology, biochemistry, chemistry, medicine, machine learning, formal ontologies, natural language processing, bioinformatics and others. It was the aim for this seminar to bring together researchers from all those areas to investigate on the state-of-the-art in both research fields, to discuss the suitability and progress of available resources, to identify areas where we are lacking tools, standards, or resources, and to foster joint opportunities for Text Mining and ontological research for the benefits of life science research.

In preparation of the seminar and prior to the meeting, the organizers identified three areas that best highlight the achievements and challenges in bringing together ontologies, Text Mining, and biological research:

(1) exploring the benefits resulting from improved relations between Text Mining and biological ontologies,

(2) technical advances in Text Mining and their application to life science research, and,

(3) success stories of Text Mining solutions with and without ontological support.

The seminar brought together more than 40 internationally renowned researchers from all domains mentioned beforehand. The ambience of the seminar is best described with the concept of a prolonged, lively and heated discussion. The discussion was mainly driven by the divergence of requirements, goals, and expectations between the Text Mining and the ontology community. On the other side, a number of talks have pointed out the successful integration of Text Mining solutions into research in ontological design and the exploitation of ontological resources for successful Text Mining solutions.

3 Ontologies for Text Mining – Text Mining for ontologies

One particularly important research question in this area is how terminological resources, such as ontologies, can best support information retrieval (IR) and information extraction (IE) solutions and vice versa. In theory we can expect that large terminological resources cover well the domain knowledge and efficiently contribute to one basic information extraction step, i.e. to named entity recognition (NER), in both IR and IE. In reality, conceptual resources such as ontologies form poor terminological resources since they have never been designed to serve this purpose. From the Text Mining perspective, they fall short to cover a significant part of the domain knowledge, i.e. they are still sparsely populated, and they do not incorporate morphological and syntactical variability (again, not the purpose of an ontological resource). On the other hand biological researchers put significant effort into the development of more and more complete ontological resources.

This topic was in the center of repeated and fruitful discussions throughout the whole meeting. Only to pick a few examples, the following presentations in the seminar showcased ongoing work spanning from Text Mining to ontologies. Paul Bruitelaar (DFKI, Saarbrücken, D) focused on the ontology life cycle, highlighting on the need to constantly adapt ontologies to the changing needs of the community. This topic was also covered by Jong C. Park (KAIST, Korea) who demonstrated his analyses on tracking changes in the gene ontology. Robert Stephens (University of Manchester, U.K.), Christopher Brewster (University of Sheffield, U.K.) and David Shotton (University of Oxford, U.K.) reported on an ongoing project for bootstrapping an ontology for animal behavior using Text Mining; initial results from similar projects for phenotype data and lipid metabolism were presented by Ulf Leser (Humboldt University, Berlin, D) and Thomas Wächter (University of Dresden, D), respectively. All three presentations highlighted the problem that despite the many papers on ontology learning, actually very few methods are readily available. Stephan Schultz (University of Freiburg, D) explained the latest developments in the BioTOP ontology which intends to bridge from top-level ontologies like BOF to domain specific ontologies such as the Gene Ontology. Studies are underway to use BioTop for word sense disambiguation, an essential step in Text Mining.

Robert Stephens pointed out in his summary of the day, that there is a certain danger that “ontologists” are disappointed by lack of perfection when using results from Text Mining for ontology development, and that Text Mining researchers are disappointed by deficiencies of existing ontologies, such as incompleteness and inadequate modeling of lexical variation in the terminology used to express the labels of the concepts. However, the seminar was successful in showing up the borderlines and the crossovers between both research domains giving inspiration to novel approaches using the best of both breeds.

4 Advanced NLP and Text Mining

Text Mining makes use of techniques from “pure”, domain-independent machine learning and natural language processing. However, many current systems in the Life Sciences use only very little linguistic information, i.e., typically only word stems or part-of-speech tags. This may lead to misinterpretations of generated evidence, since, for instance, negations and subject–object relationships are ignored. Using more linguistic information is therefore an obvious possibility to improve systems, especially as tools for generating such information in principle are available in the NLP community. However, such attempts sometimes report disappointing results. The reasons for this finding are diverse, including parsers lacking accuracy or insufficient adaptation of the extraction techniques to the representation of information in the text.

The second day of the seminar gave room to presentations on reports on technical advances in Text Mining systems and applications. Named Entity Recognition, a hot topic in the core of Text Mining for years now, was in the focus of talks by Ted Briscoe (ComputerLab, Cambridge, U.K.), Peter Murry-Rost (University of Cambridge, U.K.) and Martin Hofmann-Apitius (Fraunhofer SCAI, Bonn, D).

Ted Briscoe reported promising results on improving the accuracy of recognizing names of fly genes in text, a notoriously difficult task. The other two speakers presented latest results from applying Text Mining to chemical entities, which, in particular, include the analysis of images in text to recover chemical structures. Advances in systems for relationship extraction were presented by Goran Nenadic (University of Manchester, U.K.) and Jung-Jae Kim (EBI, Hinxton, Cambridge, U.K.). A system covering a particular important area, the resolution of anaphora in text, was shown by Su Jiang (Infocomm, Singapore). Notably, this system is also available as web service to be included in world-wide distributed Text Mining pipelines.

5 Successful Text Mining solutions

Text Mining solutions process text to enable better access, to extract well-defined results, to reduce the content to the relevant parts and, in the end, to reduce the amount of reading as the main benefit to its users. It is yet unresolved, which existing or future solution will be the best in the end. The following are some of the parameters relevant in the design of Text Mining solutions that either support improvements or, if not considered, will hinder usability: Types of data searched in the literature, types of documents available, different ways to post-process the data, interface design, linking with other resources etc. On the other hand, every successful Text Mining solution incorporates design principles, which help to understand how terminological resources and user profiles and expectations fit together.

Therefore, the third day covered talks presenting ingredients and pitfalls of successful Text Mining systems. Opportunities for getting Text Mining involved in every day curation work were explained in detail by Judith Blake (Jackson Lab), using the experience from the Mouse Genome Database as an example, including relevance classification, topic-based routing, gene name tagging and information extraction. Anna Divoli (University of Chicago, U.S.A.) presented results from two user surveys which were conducted in conjunction with the BioText project to explore on the priorities in the design of user interfaces for biological users. There was a general agreement that it is important to keep end users involved in the development phase. HM Müller (Caltech, California, U.S.A.) presented the design principles of TextPresso, which is being used by at least 20 curation teams around the world. Jörg Hakenberg and Martin Krallinger (CNIO, Madrid, Spain) reported on the development of a meta service for Text Mining tools that emerged from the second BioCreative competition, which was acknowledged as having the potential of a high impact in the field by giving access to advanced Text Mining solutions. Services were also the focus of the presentation of Dietrich Rebholz-Schuhmann, highlighting a suite of Text Mining tools hosted at the European Bioinformatics Institute. Commercial tools were presented by

Michael Schröder (GoPubMed, University of Dresden, D) and David Milward (Linguamatics, Cambridge, U.K.). An example for a very innovative application of Text Mining was shown by Nigel Collier (University of Tokyo, Jp): The BioCastor system gathers and analyses news for their relevance to indicate disease outbreaks, thus building an early warning or "rumor surveillance" system.

6 Ongoing work in the development of phenotype resources

A topic that emerged in the course of the seminar was the increasing demand and importance to manage, represent and integrate conceptual representation of phenotypes. As an immediate action, present experts in this topic reported on ongoing work and progress in this domain. Judith Blake (Jackson Laboratory, Maine, U.S.A.) presented ongoing work in the design and development of the Mammalian Phenotype Ontology at the Mouse Informatics Centre. This ontology was, among many other textual resources, used by Ulf Leser and colleagues to infer predictions of protein functions through the association of concept profiles composed of phenotypic features. Suzanna Lewis (Berkeley Drosophila Genome Project, U.S.A.) reported on the development of phenote.org, a novel resource for describing phenotype data in a very generic data format. The format reduces all representations to tuples that are formed by an ontological concept and a qualifier from a special qualifier ontology, an approach which nicely leverages existing ontologies for a new purpose. Finally, Robert Höhndorf (MPI, Leipzig, D) showed the involved logical consequences of representing "phenotypes" as derivations from a wildtype which calls for the use of non-monotonic or default logics.

7 Conclusions

The seminar brought together researchers from different research fields that are linked to Text Mining and ontologies in the life sciences and gave them a plenum to discuss their shared and disparate views. It became clear that there could be better collaborative research and that truly interdisciplinary approaches should give better results over research restricted to only one domain, but such collaborative research first of all increases the overhead and are probably not easy to sustain. It also became clear, that there are difficulties attached to collaborative work which are linked to cultural or social differences in the research work, like the question of where and what to publish to sustain individual careers. Furthermore, finding research funding for developing mature systems, ready to be used by biologists, instead of research prototypes supporting "only" a publication is difficult. This situation results in many interesting approaches that are never made available for real life applications. However, the participants clearly acknowledged that seminars such as this one are exactly the right way to overcome those problems.