Digging for Knowledge with Information Extraction: A Case Study on Human Gene-Disease Associations

Markus Bundschus
Institute for Computer Science, LMU München
Munich, Germany
bundschu@dbs.ifi.lmu.de

Anna Bauer-Mehren
Research Unit on Biomedical Informatics, IMIM, UPF
Barcelona, Spain
abauer-mehren@imim.es

Volker Tresp
Siemens AG
Corporate Technology
Munich, Germany
volker.tresp@siemens.com

Laura Furlong
Research Unit on Biomedical Informatics, IMIM, UPF
Barcelona, Spain
lfurlong@imim.es

Hans-Peter Kriegel
Institute for Computer Science, LMU München
Munich, Germany
kriegel@dbs.ifi.lmu.de

ABSTRACT

We present the information extraction system *Text2SemRel*. The system (semi-) automatically constructs knowledge bases from textual data consisting of facts about entities using semantic relations. An integral part of the system is a graph-based interactive visualization and search layer. The second contribution in this paper is the presentation of a case study on the (semi-)automatic construction of a knowledge base consisting of gene-disease associations. The resulting knowledge base, the **L**iterature-derived **H**uman **G**ene-**D**isease **N**etwork (LHGDN), is now an integral part of the Linked Life Data initiative and represents currently the largest publicly available gene-disease repository. The LHGDN is compared against several curated state of the art databases. A unique feature of the LHGDN is that the semantics of the associations constitute a wide variety of biomolecular conditions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Management, Design, Experimentation

1. INTRODUCTION

As of December 2009, there are 5414[1] journals indexed in the world's largest biomedical database PubMed. From the year 2000 to 2008 the number of articles stored in PubMed almost doubled. With each new published article, a cohort of new facts is introduced to the public. This immense growth of literature in the biomedical domain calls for automatic methods to extract these myriads of potential new findings and make this information available for search and analysis.

[1] http://www.nlm.nih.gov/bsd/num_titles.html

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

A particularly important piece of biomedical knowledge concerns genes and their association with diseases. Our knowledge about the possible genetic causes of complex diseases is still limited and for many diseases there are still no proper cures. Throughout this paper, we refer to disease genes as genes that are either involved in the causation of a disease or are associated with a disease [8]. One problem is that individual researchers are only familiar with a small portion of the available published knowledge and don't overview the literature as a whole. This phenomenon is also known as the problem of knowledge pockets [7]. As a concrete example, several gene-disease association repositories do exist on the web and they all have a special focus and thus are highly focused and non-redundant. However, many diseases are caused by the effect of several genes and thus a unified view on human gene-disease associations will help to improve the understanding of complex diseases tremendously. A particular problem is that the various databases use different controlled vocabularies, which are tailored to their specific purposes. As a consequence, data integration issues are another major challenge.

Information Extraction (IE) techniques can help to alleviate the situation. Given that all or at least most of the knowledge is somewhere available in textual form and can thus, in principle be extracted, IE approaches have a compelling property: they can alleviate the data integration problem by using consistent controlled vocabularies in advance and search subsequently through all available unstructured data sources.

1.1 Contributions and Outline

In this paper we present *Text2SemRel*, which (semi-)automatically constructs knowledge bases of entities and relations extracted from textual data. Extracted facts can be converted into a canonical form provided that controlled vocabularies of entities are available. From a semantic web perspective, the system populates an ontology of interest. *Text2SemRel* extends the relation extraction (RE) system published in [6] with an interactive visualization and search layer (see Figure 1). The system supports interactive exploration with simple keyword search over the structured knowledge base. Thus, we combine information extraction, graph-based visualization and simple keyword search into one single framework.

The second main contribution is the presentation of a case study on applying *Text2SemRel* to a large biomedical textual repository to extract semantic gene-disease associations.

The rest of the paper is organized as follows: Section 2 introduces the proposed system. The LHGDN is described in more de-

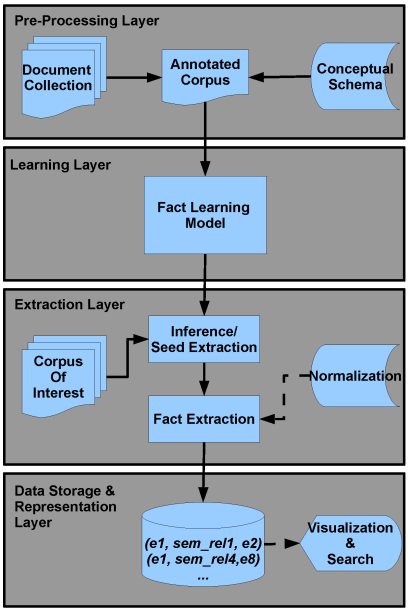

Figure 1: Text2SemRel framework. The system consists of four main layers: the pre-processing layer, the learning layer, the extraction layer and the data-storage & representation layer. The dotted lines indicate optional steps.

tail in Section 3. Before concluding in Section 5, we briefly highlight related work in Section 4. Please note that a longer version of this paper can be found online in form of a technical report[2].

2. TEXT2SEMREL

2.1 System Overview

As can be seen from Figure 1, *Text2SemRel* consists of four main layers. The first three layers are based on the relation extraction framework presented in [6].

- **Pre-Processing Layer:** The learning layer uses supervised learning, i. e. *Text2SemRel* depends on labeled training data. A corpus of documents has to be annotated with typed entities and relations. As a consequence it is more suitable for targeted relation extraction. Entity classes and relations between entity classes originating from a given conceptual schema or semantic data model are aligned with a document collection of interest in the pre-processing phase. In case the document collection refers to a wikipedia-style article collection, where a so-called key entity is given in advance, the relation extraction problem is posed slightly differently (see Section 3, 2nd paragraph). We argue that *Text2SemRel* is quite general and can easily be extended to a new domain provided that labeled training data is available. As a concrete example, *Text2SemRel* was also applied to extract semantic relations between diseases and treatments in our previous work [6].

- **Learning Layer:** The approach is based on Conditional Random Fields (CRFs). The fact learning module comprises two challenges: Learning feature weights for extracting named entities from text (Named Entity Recognition (NER)) and learning feature weights for extracting typed relations between them (referred to as Semantic Relation Extraction (SRE)). Hereby, we cast the problem of NER and SRE into a sequence labeling problem. Due to their discriminative nature, CRFs can easily incorporate arbitrary, non-local features from the input sequence. This characteristic will be very suitable for tackling the task of semantic relation extraction. In particular, we use three main types of features: local, context and external knowledge features. See [6] for more details.

- **Extraction Layer:** During inference, we use the well-known Viterbi algorithm to label unannotated texts. Afterwards we convert the labeled text to facts. To follow the *Linked Data principles* [5], facts can be normalized in our system in order to exploit the full power of the gained structured data. Using already existing URI's is the preferred way to publish Linked Data. Thus, if a controlled vocabulary for entities of the modeled domain of interest is already available, this vocabulary can be used as reference for normalization. A simple sliding-window heuristic is currently used in this step. Entity mentions recognized by the CRF are compared to a dictionary. The textual phrases of the entity mentions are expanded or reduced, until the most specific controlled vocabulary entry has been found. Note that the relations come already in normalized form, since the system is tackling targeted relation extraction.

- **Data-Storage & Representation Layer:** The extracted facts are encoded in form of subject, predicate and object triples, using the Resource Description Framework (RDF). In our case, subjects and objects are typed entities and the predicate encodes a relation that holds between the involved entities. As an example, the fact that the expression of gene ITGB4 is altered in thyroid carcinoma, is encoded with the following triple:
  <http://bio2rdf.org/geneid:3691, http://www.anonymous.de/AlteredExpression, http://bio2rdf.org/mesh:D009362>
  Optionally, additional information such as the publication ID or the publication itself can be stored as well. This is solved by means of reification.

2.2 Visualization and Interactive Search

Once a knowledge base consisting of triples has been constructed with *Text2SemRel*, the question arises of how it can be accessed. Besides common filtering functions over the knowledge base such as filtering for specific entities and relations, *Text2SemRel* comes with an easy to use, graph-based visualization framework. The system supports interactive exploration with simple keyword search over the Entity-Relationship (ER) graph. Thus, we use a combination of paradigms originating from two different communities to access knowledge bases derived from text: (i) graph-based visualization and (ii) keyword search over structured data. The innovative combination of the just mentioned paradigms for information access in *Text2SemRel* makes extracted knowledge bases easily searchable.

Graph-based Visualization.

The underlying data structure of *Text2SemRel* is an ER graph, consisting of entities such as genes, drugs, diseases etc. and of relations concerning the entities. Thus, visualizing the knowledge base as a graph is intuitive and enables the user to get a fast overview. E. g. , as a benefit of the graph-based visualization it can be easily seen which entities are the most connected ones (the hubs) in the ER-graph. Moreover, it can easily be seen which parts of the

[2] www.dbs.ifi.lmu.de/~bundschu

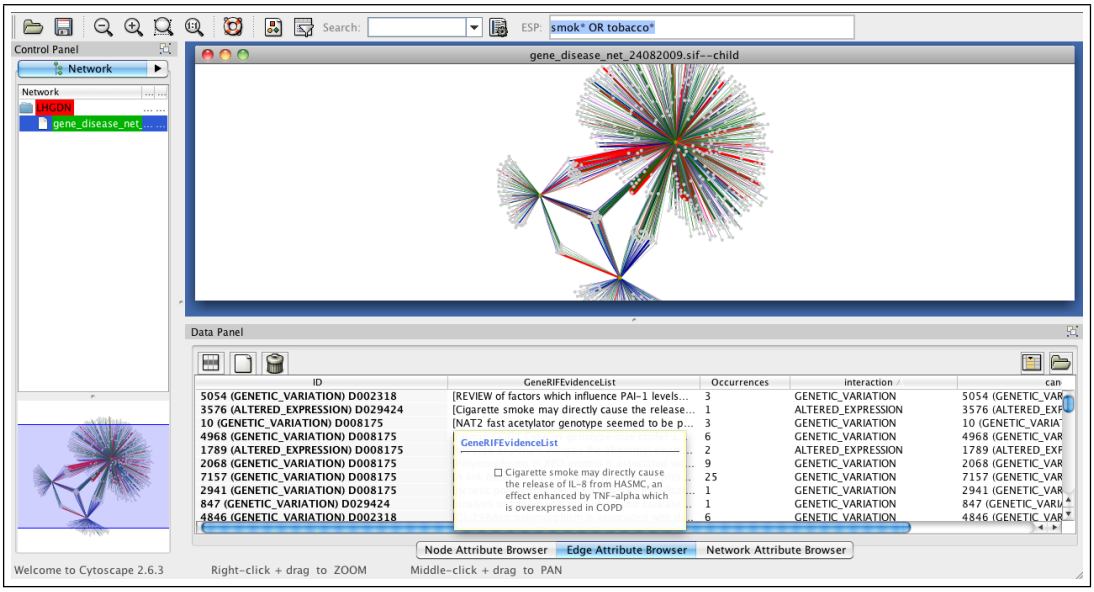

Figure 2: Screenshot of a subnetwork that shows all facts for three diseases. The subnetwork of COPD, CVD and lung cancer are shown. Green squares are diseases, while the grey dots represent the entity type gene. The color of the edges indicates the type of relation as predicted by *Text2SemRel*. Edges highlighted in red indicate that the keyword query *smok* OR tobacco are matched. The thickness of the edges indicate the number of times a fact was found in different publications. The data panel provides important additional context information such as the sentence from which the fact was extracted.**

knowledge base are disconnected. All these things make the graph-based visualization a helpful feature for knowledge discovery.

Keyword Search over Structured Data.

This second component for simplifying information access in *Text2SemRel* is motivated by the attempt to make relational database management systems (RDBMS) as easily searchable as keyword-based search engines (see e. g. [1]). A common criticism of formal query languages is that it is hard and uncomfortable for end-users to pose queries. Furthermore, despite the recent advances in the IE domain, it will not always be possible to convert all relevant information into a structured form. One reason for this is that the assessment of relevance regarding information is highly subjective and varies from end-user to end-user. Therefore, methods are needed that are able to extract the most important facts (semi-) automatically, but at the same time are able to further narrow down the desired context. *Text2SemRel* makes use of the plain text collections from which the facts are extracted and annotates the facts with words originating from these sources. Put in other words, a fact is treated as a bag-of-words vector consisting of all words occurring from the various source documents from which the fact was extracted. In this way, we can additionally filter facts according to Boolean keyword queries and thus provide additional powerful filtering capabilities, without the need to extract further entities and relations. Thus, the keyword-search component over the structured ER-graph will help to satisfy the manifold end-user needs.

System Description.

The starting point of the visualization is the complete knowledge base. A user has the following options to search for specific facts:

- Filtering for entities and/or types of relations and/or attributes such as in traditional RDBMS systems.

- Filtering of facts with regards to keywords. We index all available information for facts by using the enhanced search plugin (ESP) [2]. The ESP allows advanced Boolean querying such as search with wildcards or range search for numbers. One of the unique features of a literature-derived network is that we can use the available unstructured information for filtering purposes. To keep sensitivity high, we only include the words of the sentence from which the fact was extracted.

- Merging of networks. Every filtering step induces a new subnetwork. These subnetworks can be merged to investigate connections between different subnetworks. Thus, we can easily isolate information but also assemble disconnected information.

The information access layer in our system employs Cytoscape[2], which is a network visualization tool well established in the biomedical community. Furthermore, the Boolean keyword query language is implemented with the ESP, that uses the Lucene retrieval library[3] for indexing. Note that a complex retrieval example is described in the longer technical report[2].

3. LHGDN

Text2SemRel was trained on an in-house generated training corpus (see [6] for more information). In contrast to the gene-disease associations provided by the curated state of the art databases, gene-disease associations in the LHGDN are classified into several biomolecular conditions. These conditions are describing a wide variety of molecular events, ranging from genetic to transcriptional and phosphorylation events. In particular, we defined the following conditions that can hold between genes and diseases: *altered expression, genetic variation, regulatory modification, any relation and negative associations*. It was shown in former experiments that the system is able to extract the just described gene-disease associations with a F-measure of 78% (NER+SRE) (see [6]).

[2] http://www.cytoscape.org/features2.php
[3] http://www.lucene.apache.org/

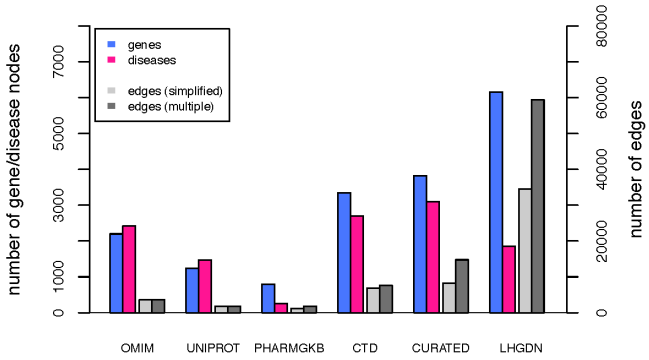

Figure 3: Number of genes, diseases and edges in the databases under consideration. "Edges simplified" represent unique gene-disease associations, while "edges multiple" denotes that an association is counted every time it is mentioned in a database.

Text2SemRel was applied to the whole Entrez Gene's[4] GeneRIF (Gene Reference Into Function) database. The GeneRIF database represents a rapidly growing knowledge repository and consists of high-quality phrases created or reviewed by Medical Subject Heading (MeSH) indexers. Hereby, the phrases refer to a particular gene in the Entrez Gene database and describe its function in a concise phrase. The LHGDN was created based on a GeneRIF version from March 31st, 2009, consisting of 414241 phrases. These phrases were further restricted to the organism Homo Sapiens, which resulted in a total of 178004 phrases. In the GeneRIF database the genes are already given in advance and thus the relation extraction problem converts to (i) identifying disease mentions and (ii) deciding which relation holds between the given key entity (here a gene) and the other entities in the sentence. Due to the particular structure of the GeneRIF database, the gene normalization is trivial, which otherwise is in itself a serious research problem (see e. g. the BioCreAtIvE[5] evaluations). The identified disease mentions were normalized with a simple sliding-window heuristic to Bio2RDF[6] URI's.

Figure 3 compares the size of the LHGDN with several other state of the art databases[7][8][9][10]. CURATED represents a current integration effort of several manually curated databases [4]. It can be easily seen, that a tremendous amount of knowledge about gene-disease associations is currently locked in the literature. Besides this, the LHGDN was compared to the curated databases with respect to biological properties of the genes, in particular pathway and GO (Gene Ontology) homogeneity. Furthermore, we investigated carefully the distribution of gene, disease and fact mentions in the LHGDN. It turns out that all three empirical quantities in the current version of the LHGDN follow a power-law distribution. This gives interesting insights in the underlying processes of how facts centered around gene-disease associations are published. Please refer to the technical report for a detailed description of the results[2].

[4] http://www.ncbi.nlm.nih.gov/gene
[5] http://www.biocreative.sourceforge.net
[6] http://www.bio2rdf.org
[7] http://www.ncbi.nlm.nih.gov/omim
[8] http://www.uniprot.org/
[9] http://www.pharmgkb.org/
[10] http://ctd.mdibl.org/

4. RELATED WORK

[9] provides a recent survey over the broad field of IE and aligns existing work with several dimensions. *Text2SemRel* treats the task of Semantic Relation Extraction as sequence labeling task. Most similar to our approach in the algorithmic sense is the *TextRunner* system [3]. However, *TextRunner* aims at Open Information Extraction, a very ambitious, relation-independent extraction paradigm. Thus, the system may not yield high precision at reasonable recall when compared to traditional RE frameworks [3]. Another system using CRF-based technology for extracting relations is the Kylin system for generating Wikipedia infoboxes [10].

5. CONCLUSION

With *Text2SemRel* we introduced a framework that is able to extract facts from textual resources and thus contributes to reduce the gap between text and knowledge. *Text2SemRel* allows expressive search and interactive exploration over the extracted knowledge base and thus facilitates knowledge discovery. As a result from applying *Text2SemRel* to a large biomedical text collection, we presented the LHGDN, which currently is the largest gene-disease association repository publicly available. The LHGDN is an integral part of the Linked Life Data[11] initiative, which confirms the high quality of facts extracted with our presented system. Last but not least, we investigated how entity and fact mentions are distributed in the LHGDN. A careful statistical analysis of the current snapshot of the LHGDN reveals that the number of entity and fact mentions follows a power law distribution.

6. REFERENCES

[1] Agrawal et al. Dbxplorer: A system for keyword-based search over relational databases. In *ICDE '02*, Washington, DC, USA, 2002.

[2] M. Ashkenazi et al. Cytoscape esp: simple search of complex biological networks. *Bioinformatics*, 24(12):1465–1466, June 2008.

[3] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08*, pages 28–36, Columbus, Ohio, June 2008.

[4] Bauer-Mehren et al. Network analysis of an integrated gene-disease association database reveals functional modules in human disease. *Submitted*, 2010.

[5] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[6] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H. P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207+, 2008.

[7] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky. Emergent behavior of growing knowledge about molecular interactions. *Nature Biotechnology*, 23(10):1243–1247, October 2005.

[8] M. G. Kann. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Brief Bioinform*, 11(1):96–110, January 2010.

[9] S. Sarawagi. Information extraction. *Found. Trends databases*, 1(3):261–377, 2008.

[10] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In *CIKM '07*. ACM, 2007.

[11] http://www.linkedlifedata.com