KnowLife: a Knowledge Graph for Health and Life Sciences

Patrick Ernst [#1], Cynthia Meng [*2], Amy Siu [#3], Gerhard Weikum [#4]

[#] *Databases and Information Systems, Max-Planck Institute for Informatics*
Campus E1 4, 66123 Saarbrücken, Germany
[1,3,4] {pernst, siu, weikum}@mpi-inf.mpg.de

[*] *Harvard School of Engineering and Applied Sciences*
29 Oxford Street, Cambridge, MA 02138, USA
[2] cynthiameng@college.harvard.edu

***Abstract*—Knowledge bases (KB's) contribute to advances in semantic search, Web analytics, and smart recommendations. Their coverage of domain-specific knowledge is limited, though. This demo presents the KnowLife portal, a large KB for health and life sciences, automatically constructed from Web sources. Prior work on biomedical ontologies has focused on molecular biology: genes, proteins, and pathways. In contrast, KnowLife is a one-stop portal for a much wider range of relations about diseases, symptoms, causes, risk factors, drugs, side effects, and more. Moreover, while most prior work relies on manually curated sources as input, the KnowLife system taps into scientific literature as well as online communities. KnowLife uses advanced information extraction methods to populate the relations in the KB. This way, it learns patterns for relations, which are in turn used to semantically annotate newly seen documents, thus aiding users in "speed-reading". We demonstrate the value of the KnowLife KB by various use-cases, supporting both layman and professional users.**

I. Introduction

Motivation: Knowledge bases (KB's) like dbpedia.org, freebase.com, and yago-knowledge.org have become a great asset in interpreting and enriching Web contents for entity-relationship-oriented search, recommendations, and analytics [1], [2]. The Google Knowledge Graph and the IBM Watson technology are prominent examples. Projects of this kind have looked at entities and facts in a broad, general-purpose manner. However, interesting use-cases often require domain-specific knowledge at a depth and coverage that universal KB's do not provide. The domain that we focus on is health and life sciences.

Databases on the Web contain a wealth of information about proteins, genes, and molecular pathways, and there is also an enormous amount of health-oriented, textual information on diseases and drugs available in specialized portals and discussion forums. Besides scientific publications found in PubMed Medline, physicians as well as laymen also consult health portals on the Web, such as uptodate.com or mayoclinic.com/health-information. Moreover, there are rapidly growing online communities, such as www.patient.co.uk, sharing experience and knowledge about health issues, such as side effects of drugs and drug combinations, or symptoms of diseases.

All this constitutes a rich universe of health information, but the information is spread across many sources, mostly in textual form, and unorganized – far from being anywhere near a semantic KB. Our research presented here aims to fill this gap by automatically constructing and maintaining a KB.

Prior Work and its Limitations: Biomedical KB's, such as disease-ontology.org, omim.org, or drugs.com, are manually built and curated and specialized on a single aspect like genetically inherited diseases, human anatomy, or FDA-approved drugs. It is very tedious to inter-connect all these sources in a clean and informative way. What is missing is a *one-stop portal* that comprises knowledge on all health-related aspects in an integrated manner.

Research on information extraction (IE), e.g., [3], [4], [5], [6], [7], aims to extract life-science-oriented relational facts from natural-language text sources such as PubMed publications. State-of-the-art work has two major limitations:

1. It mostly focuses on the molecular level; a typical IE task is to extract protein-protein interactions.
2. It solely taps into scientific literature, and disregards social media like Web portals and discussion forums.

There is increasing awareness in life sciences that a more holistic view is needed: for example, relating genetic factors of diseases to other risk factors such as nutritional habits and life style, looking into side effects of combinations of drugs rather than single drugs alone, or analyzing the experience of patients on mass diseases such as asthma or diabetes.

There is little prior work that pursues a holistic kind of IE and KB construction. [8] adapts the NELL framework for learning relational extraction [9] to the biomedical domain. [10] harnesses the IE methods of [11] for populating a broad range of disease-centric relations, including symptoms, anatomic parts, environmental and life-style risk factors, etc. Unlike most IE work, this approach maps all arguments of relations into canonicalized (i.e., uniquely identified) entities registered in KB's, and uses logical consistency reasoning to achieve near-human precision.

Contributions: The project demonstrated in this paper, called *KnowLife*, builds on the preliminary work of [10], but extends

it in various ways and constructs a much richer one-stop KB. In contrast to prior work, we tap into both life-science publications and health-related online forums, and integrate the extracted facts with biomedical backbone knowledge. All facts in KnowLife carry provenance information, so that users can explore the evidence for a relational fact. The acquired knowledge, including textual patterns for relations, is used to annotate any kind of input document, expert-level or layman style, with entities and relationships on the fly, as the user reads it. The value of the KnowLife portal is demonstrated by several use-case scenarios: laymen exploring health issues of personal interest, medical professionals searching for specific knowledge, and researchers "speed-reading" publications via entity-relationship synopses.

The salient contributions of the KnowLife project can be highlighted as follows:

- Using advanced IE methods for constructing a large KB on a wide range of health-centric relations with entity linking to the Unified Medical Language System (UMLS, `uts.nlm.nih.gov`): a total of 214k canonical entities and 78k facts for 14 relations.
- Tapping into health-related online forums for KB population and for providing evidence for relational facts.
- Automatically annotating newly seen documents from scientific literature or from social media with relevant entities and relationships mentioned in natural-language form.

The KnowLife portal is available for interactive use at `https://gate.d5.mpi-inf.mpg.de/knowlife`.

II. Knowledge Harvesting from Web Sources

Our system extracts crisp facts from textual Web sources. We support 14 relations with semantic type signatures, for example, $isSymptom : symptom \times disease$, $createsRisk : riskFactor \times (disease \cup symptom)$, $treats : drug \times (disease \cup symptom)$, $observedIn : (disease \cup symptom) \times bodyPart$, etc. These relations hold between entities of specific types. To this end, we build on the entity thesaurus provided by UMLS, which contains several million entities in total. Our approach to information extraction from Web documents combines statistics-based pattern matching for high recall of fact candidates with logics-based consistency reasoning for high precision of eventually accepted facts [11]. This two-stage process is driven by a small set of manually compiled seeds. For this purpose, we use 1026 seed facts extracted from 52 `uptodate.com` articles. We tap into two kinds of input sources: scientific publications and postings in health portals. Examples for the typical sentences that we aim to tap are:
"High-dose cyclobenzaprine resulted in a significant reduction in the Tinnitus Handicap Inventory (THI) score between baseline and week 12 in the intention-to-treat sample." (from the abstract of a scientific paper in Pubmed), and
"This rare type of tinnitus may be caused by a blood vessel problem, an inner ear bone condition or muscle contractions." (from the patient-oriented portal of the Mayo Clinic).

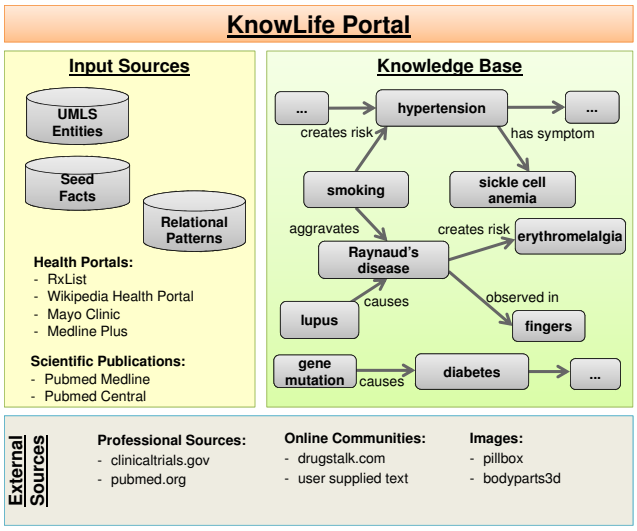

Fig. 1: Overview of the KnowLife Portal

For building the KnowLife portal, we processed 593,423 abstracts from Pubmed Medline and Pubmed Central, and 28,650 Web pages of the following portals: `rxlist.com`, `mayoclinic.com/health-information`, `wikipedia.org/wiki/Portal:Health`, and `nlm.nih.gov/medlineplus`. We process these inputs in three stages:

1. *Named Entity Recognition and Disambiguation:* We use our own dictionary-based method [12] to identify sentences in Web sources that may potentially express a relational fact. From the six Web sources stated above, we extracted a total of about 6.5 million sentences, giving us noisy fact candidates. The entity names in these sentences are mapped to UMLS. Where multiple entities are mapped to the same text occurrence, we use the type hierarchy of UMLS to select the most specifically typed entity.
2. *Pattern Mining:* We use the Prospera tool [11] for computing salient patterns that connect the entities in the candidate sentences. This step is based on a frequent-itemset-mining algorithm. The patterns are weighted by statistical analysis, in terms of confidence and support. Here, we use the seed facts and their co-occurrences with certain patterns as a basis to compute confidence. The best patterns, with both confidence and support above specified thresholds, serve to filter the initial set of candidate facts, from 6.5 million down to 212k facts. The patterns are also stored in the KB, as they can later be used for annotating relationships between entities in newly seen documents that a portal user wants to "speed-read". In total, we identified 178 salient patterns for the 14 relations in the KB.
3. *Consistency Reasoning:* We use a Weighted Max Sat Solver to reason over the mutual consistency of the fact candidates, and accept a consistent subset with a high total weight (ideally the subset with the maximum total weight, but we obviously use an approximation algorithm due to the NP-hardness of the problem). This way, we accepted about 78k high-quality facts. Typical consistency

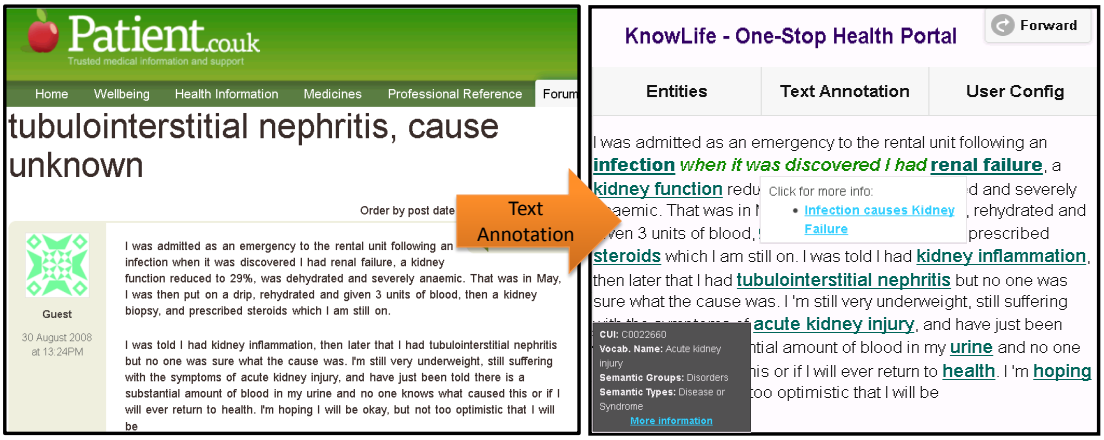

Fig. 2: KnowLife text annotation of an excerpt from `www.patient.co.uk`

constraints leveraged here are the type signatures of relations and the mutual exclusion between certain pairs of relations. For example, if a drug has a certain symptom as a side effect, this rules out that the drug treats this symptom.

As an extension to the facts in its KB, KnowLife can import textual evidence for facts from the input sources, and additional information about entities or facts from external sources. Here, we tap into additional portals, for example, to obtain images of anatomic parts, or discussion threads from social media. These external sources need to be registered in KnowLife. Once this is done, we use their query interfaces to retrieve relevant matches for an entity or two entities in a fact. Strictly speaking, there may be a name-to-entity disambiguation problem here, but we use the canoncial names for querying and almost always retrieve correct results. Example sources that we tap into are `lifesciencedb.jp/bp3d` (for anatomic images), `clinicaltrials.gov`, `drugtalk.com`, `patient.co.uk`, and more. Figure 1 gives an overview of the KnowLife architecture.

III. Text Annotation

In addition to harvesting knowledge from a large collection of textual sources, KnowLife automatically annotates ad-hoc text on the fly. This allows a user to "speed-read", as the system applies already acquired knowledge to newly seen documents, annotating entities and facts in real time. This functionality is a major extension that distinguishes KnowLife from other biomedical KB's. The user can copy-and-paste text or provide a URL for a Web page of interest. In the latter case, we use the CETR tool [13] for removing boilerplate information and casting the HTML input into plain text.

The on-the-fly annotation uses the following two steps:

1. *Pattern Matching:* We use a dictionary-based method for discovering entities and mapping them to the KB (see Section II). Wherever two entities co-occur in a sentence, we use our repertoire of salient patterns for relations, learned during the knowledge harvesting. We match each pattern against the sentence and collect partial-match results. The results are scored based on their word-level Jaccard overlap with the relation pattern. Patterns with a score above a specified threshold are considered as fact candidates. As an example, consider the excerpt of an online discussion from `www.patient.co.uk`, shown in Figure 2. The text "when it was discovered I had" matches 7 salient patterns belonging to 2 different relations: $causes$ and $isSymptomOf$. Thus, this text becomes a fact candidates for these 2 relations.

   This pattern matching procedure is efficient for real-time responses, and its decision power is based on previously learned, high-confidence patterns. In contrast to the more elaborate pattern analysis for the knowledge harvesting stage, we avoid expensive computations.
2. *Type Checking:* Using the type signatures of the relations, we can further filter out fact candidates whose arguments have types that are not compatible with the relation. The remaining candidates are accepted as facts and marked in the ad-hoc input text. Returning to our example, the 7 fact candidates are now whittled down to 1 fact: $infection\ causes\ renalFailure$. For the other candidate relation, $isSymptomOf$, the type checking prunes this interpretation, as $infection$ and $renalFailure$ are of both type $disease$, whereas $isSymptomOf$ would expect a left-hand argument of type $symptom$.

IV. Demo Scenarios

The KnowLife portal can be searched and browsed in many ways, supporting both laymen and professionals in knowledge discovery. Figure 3 shows a screenshot of KnowLife after retrieving an entity, the drug diclofenac. The capability for on-the-fly text annotation has already been shown in Figure 2.

Conference participants will be able to interactively query and explore the rich contents of KnowLife, and will also get a deeper understanding of the underyling information extraction and text annotation capabilities. KnowLife can also be accessed with its full functionality on the iPad.

Below, we discuss two use-case scenarios to illustrate the benefits of KnowLife for different kinds of end-users.

Layman Scenario: Consider the patient who wrote the post in Figure 2. After she was told that she has either kidney inflammation or tubulointerstitial nephritis, the diagno-

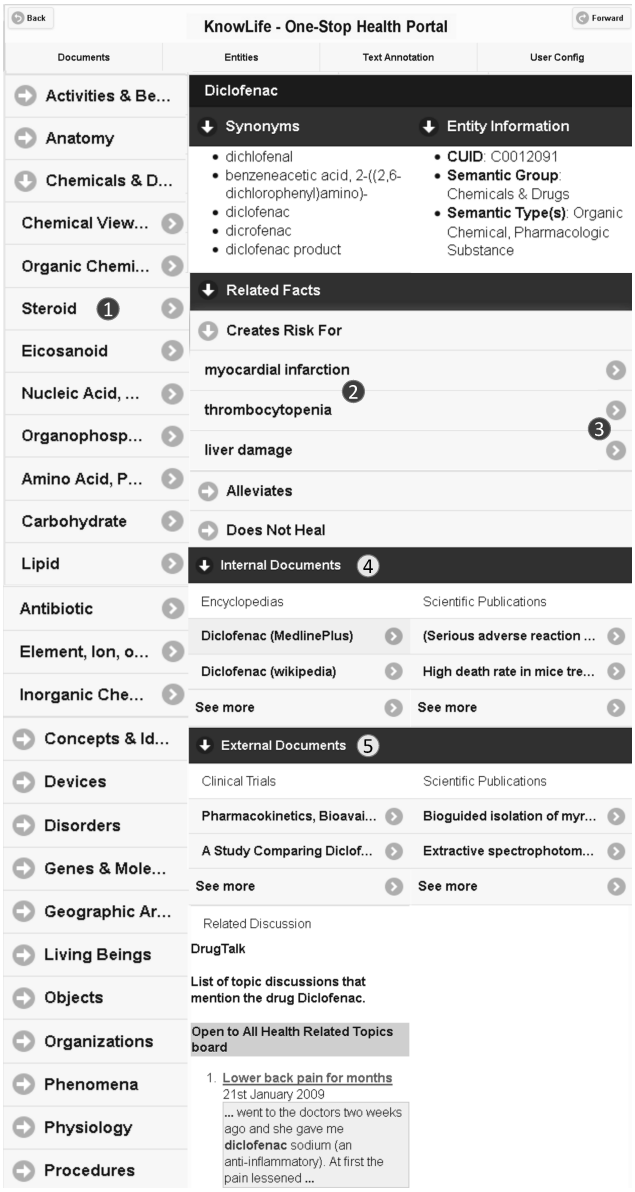

Fig. 3: KnowLife portal for exploring entities: The user can browse over the entity hierarchy (1), jump to related facts (2) or to their textual evidence (3). Relevant internal documents (4) and documents pulled from external sources (5) are listed as well.

sis remains unclear. Instead of tediously reading through many health portals, the patient can annotate her own post in KnowLife. This allows her to quickly gain access to relevant information about the recognized entities, as their links springboard her to info boxes detailing causes, risk factors, symptoms, and more. For instance, she opens the info box for *acute kidney injury* and notices that *diclofenac*, a drug she has taken in the past, causes acute kidney injury. Therefore, she can explore textual evidence for the fact $acuteKidneyInjury\ causedBy\ diclofenac$. In this case, the system leads her to a Wikipedia article prepared for speed-reading via annotated entities and facts (see Figure 4). Alternatively, she can look at the info box of diclofenac shown in Figure 3. Here, she can benefit from the experiences of other patients taking this drug by visiting the related discussions

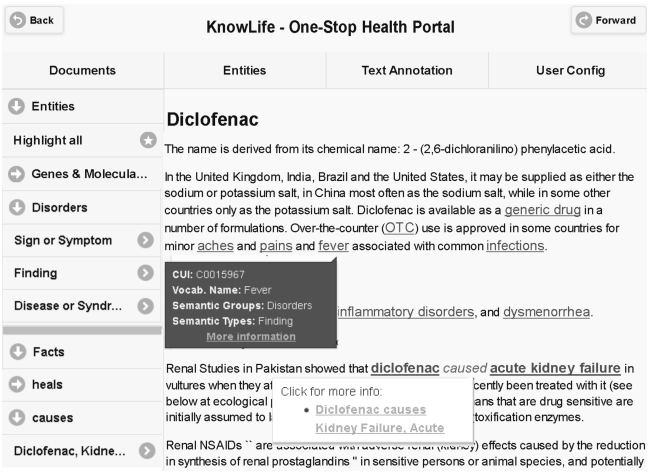

Fig. 4: Document about diclofenac annotated for speed-reading

section.

Health Care Professional Scenario: Physicians and other medical professionals are often more interested in scientific literature, so as to deepen their knowledge in specific areas and stay up-to-date with the latest research. Consider lupus, a challenging disease to diagnose and treat, since there is no standard test and no standard treatment for it. The physician's goal is to quickly obtain relevant scientific information aggregated from multiple articles. To aid diagnosing lupus, the info box lists genes that can be risk factors. By following the links in the info box, the professional can quickly jump to the latest scientific publications. For treating complex lupus cases, it is also necessary to know about clinical trials, which are directly accessible from the info box.

REFERENCES

[1] D. Barbosa, H. Wang, C. Yu: Shallow Information Extraction for the Knowledge Web. Tutorial, ICDE 2013
[2] F. Suchanek, G. Weikum: Knowledge Harvesting from Text and Web Sources. Tutorial, ICDE 2013
[3] B. Rosario and M. A. Hearst: Classifying Semantic Relations in Bio-Science Texts. ACL 2004
[4] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.-P. Kriegel: Extraction of Semantic Biomedical Relations from Text using Conditional Random Fields. BMC Bioinformatics, 9:207, 2008
[5] P. Palaga, L. Nguyen, U. Leser, J. Hakenberg: High-Performance Information Extraction with AliBaba. EDBT 2009
[6] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. Tsujii, S. Ananiadou: Event Extraction across Multiple Levels of Biological Organization. Bioinformatics 28:575–581, 2012
[7] M. Krallinger, A. Valencia, L. Hirschman: Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology. Genome Biology 9, 2008
[8] D. Movshovitz-Attias, W.W. Cohen: Bootstrapping Biomedical Ontologies for Scientific Text using NELL. BioNLP Workshop 2012
[9] A. Carlson, J. Betteridge, R.C. Wang, E.R. Hruschka Jr., T.M. Mitchell: Coupled Semi-Supervised Learning for Information Extraction. WSDM 2010
[10] V. N. Romero, M. Ye, M. Albrecht, J.-H. Eom, G. Weikum: DIDO: a Disease-Determinants Ontology from Web Sources. WWW 2011
[11] N. Nakashole, M. Theobald, G. Weikum: Scalable Knowledge Harvesting with High Precision and High Recall. WSDM 2011
[12] A. Siu, D. B. Nguyen, G. Weikum: Fast Entity Recognition in Biomedical Text. KDD Workshop on Data Mining for Healthcare 2013
[13] T. Weninger, W.H. Hsu, J. Han: CETR: Content Extraction via Tag Ratios. WWW 2010