Connecting Science Data Using Semantics and Information Extraction

Evan W. Patton and Deborah L. McGuinness

Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180 USA
{pattoe, dlm}@cs.rpi.edu

Abstract. We are developing prototypes that explicate our vision of connecting personal medical data to scientific literature as well as to emerging grey literature (e.g., community forums) to help people find and understand information relevant to complex medical journeys. We focus on robust combinations of natural language processing along with linked data and knowledge representation to build knowledge graphs that help people make sense of current conditions and enable new manners of scientific hypothesis generation. We present our work in the context of a breast cancer use case. We discuss the benefits of biomedical linked data resources and describe some potential assistive technology for navigating rich, diverse medical content.

Keywords: knowledge representation, explanation, clinical notes, natural language, web forums, nanopublications

1 Introduction

As scientific knowledge continues to grow in size and diversity, it is increasingly difficult to discover and manage information relevant to any particular context. It can be challenging to determine how a statement or report relates to others and to form and evaluate (often competing) hypotheses, e.g. related to diagnosis or treatment paths. Complications grow when content is both structured and unstructured, and when some is from less accredited sources. We aim to expand the boundaries of Linked Science by focusing on evidence modeling from natural language processing techniques (NLP) over broad content and by identifying promising data-driven hypotheses using linked data and nanopublication style encodings. We present this discussion in the context of a breast cancer demonstration use case informed by challenges experienced during a co-author's recent cancer journey. Cancer is a complex disease to manage and treat, often requiring chemotherapy, surgery, radiation, and drugs to reduce recurrence. We show how management of this information by the patient is aided by semantic technologies combined with natural language processing algorithms.

> A breast cancer patient wishes to better understand her diagnosis and planned treatment. She is interested in expected chemotherapy side effects, and leveraging experiences of other similar individuals to proactively find and evaluate promising coping strategies. She reads through

> oncologist-provided documents about her proposed chemotherapy drugs and uses search engines to find more about likely adverse effects that appear detrimental to her quality of life. She finds conflicting opinions on the efficacy of different coping strategies, and needs to determine an approach to effectively weigh the possible pros and cons. Managing this information is mentally taxing and can easily overwhelm a patient.

Our patient needs to find and comprehend potentially conflicting evidence about treatment options and side effects. We propose new software, using a variety of artificial intelligence tools built on the interoperability principles promulgated by linked data and the Semantic Web, to address these challenges.

2 Evidence Modeling

The patient uses current technologies to obtain information about her treatment strategy and to formulate promising side effect mitigations. This can be time consuming for anyone, but more so for medically naïve patients. Furthermore, technologies such as web forums or social networking sites are becoming increasingly common for discourse between patients as they can often include anecdotal reports, that have not yet been validated through clinical trials, but may be valuable. They are often presented in layperson terms and sometimes attract new patients who may be less medically literate. Due to lack of scientific rigor, there may be contradictory or unsupported information available, as shown in the following two answers about a mitigation for the very common, taxol-related, nail bed problem:

> My onc[ology] nurse told me to rub tea tree oil into my cuticles and nails every night. It is a natural anti-septic and for whatever reason can sometimes help prevent nail infections and lifting during taxol. [1]

> I wouldn't use tea tree oil. A friend did on some cracked skin and it got worse. [2]

The first suggestion is a common preventive approach for nail problems: tea tree oil prevents nail infections because "it is a natural anti-septic" and appeals to authority "my onc nurse told me to...". The second suggestion from a different user in the same thread advises against tea tree oil as "a friend [applied tea tree oil] on some cracked skin and it got worse." Natural Language techniques may be used to extract coping strategies for particular conditions but without deeper knowledge, provenance, and tools, the user may not know how to evaluate and/or integrate potentially contradictory suggestions. We are extending joint extraction techiques proposed in [4] with semantic background knowledge to aid in extracting linked data from medical records.

[1] https://community.breastcancer.org/forum/69/topic/783573
[2] https://community.breastcancer.org/forum/96/topic/745475

3 Hypothesis generation using Nanopublications

The Repurposing Drugs using Semantics (ReDrugS) project [5] has focused on modeling evidence using small units of publishable information called Nanopublications [2]. ReDrugS utilizes linked data sources to build a knowledge base of nanopublications that is then reasoned about using probabilistic techniques to identify potential links between proteins, drugs, binding sites, and genes, with the ultimate aim of discovering possible new off-label uses for FDA-approved drugs. This project's success has been partially due to the large corpus of linked data and ontologies generated by the biomedical community over the past few decades. ReDrugS has ingested content from 17 structured curated data sources, including content concerning drugs, alternate names, conditions, and pathways. Once a chemotherapy protocol is extracted from medical notes, ReDrugs can be used to find alternative drug names along with related conditions. This framework, along with the side effect resource SIDER in process, can be used to improve the patient's process in finding chemotherapy drug side effects and some mitigations by applying its search techniques to authoritative drug resources, such as looking for anti-nausea prescription drugs. The infrastructure for this system could be repurposed for other scientific domains, but only if linked data sources are abundant in those domains or if quality linked data can be generated from automated methods, e.g. via natural language processing of web-based resources.

4 Explanations

We aim to provide extensive explanation mechanisms since explanation is a key component of transparent systems and user studies have shown that explanations are required if agents are to be trusted [1]. We aid explanation generation through the collection of provenance, modeled using the W3C's PROV ontology [3]. PROV-O is a standard for modeling provenance information on the web, which allows tools to integrate distributed provenance information from different systems. We use this provenance to help construct end user explanations that include both lineage of content and support (and opposition) for a statement.

We identify potential evidence on the use of tea tree oil in chemotherapy-induced nail bed problems. Not only would a patient want to know evidence, source, and authoritativeness for both views, she might also want the system further decompose these arguments and present supporting evidence as to the antimicrobial nature of tea tree oil in more authoritative sources (e.g. [6]).

We claim that we can reuse the ReDrugS content to find prescription drugs for chemotherapy side effects. Provenance may be displayed to show that the recommendation is from a validated authoritative source. While that framework was originally designed to find potential new off-label uses for drugs along with confidence ratings, the explanation component is more critical for our use so that researchers may inspect evidence sources and the methods used to determine the system confidence. Without such explanations, people would have difficulty evaluating competing suggestions.

Our systems[3] provide explanation drill down so users can obtain as much detail as they desire, thus allowing a patient to find, for example, if authoritative sources contain prescription drugs for coping with a particular side effect. Our NL-based extraction work can be used to identify alternative, possibly competing, therapies, e.g. an herbal remedy recommended anecdotally with potentially corroborating authoritative sources.

5 Discussion and Summary

Natural Language Processing can expose some of the unstructured content of medical records as structured content as well as assist in generating linked data from unstructured sources. The ReDrugS framework provides a semantically-integrated system combining many different structured biomedical resources to generate a broadly reusable knowledge graph. By integrating the natural language and structured knowledge representation approaches, we can obtain a much richer annotated knowledge base that includes source and confidence information. Our prototypes demonstrate some ways that this rich resource may then be used to help patients and their support networks to discover, integrate, and evaluate information relevant to complicated medical situations and to help form transparent and data-driven hypotheses about how to proceed. We believe these efforts demonstrate some opportunities for future AI-enhanced Linked Science-based assistants that use the wealth of structured content as well as the growing grey literature collection.

Acknowledgements

The authors thank Heng Ji and Alex Borgida for their discussions that helped shape this work.

References

1. Glass, A., McGuinness, D.L., Wolverton, M.: Toward establishing trust in adaptive agents. In: 13th Intl Conference on Intelligent User Interfaces. pp. 227–236 (2008)
2. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use 30, 51–56 (2010)
3. Lebo, T., Sahoo, S., McGuinness, D.L.: PROV-O: The PROV ontology. Tech. rep., W3C (2013)
4. Li, Q., Ji, H.: Incremental joint extraction of entity mentions and relations. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)
5. McCusker, J., Solanki, K., Chang, C., Dumontier, M., Dordick, J., McGuinness, D.L.: A nanopublication framework for systems biology and drug repurposing. In: CSHALS 2014 (2014)
6. Pazyar, N., Yaghoobi, R., Bagherani, N., Kaerouni, A.: A review of applications of tea tree oil in dermatology. International Journal of Dermatology pp. 784–90 (2013)

[3] http://tw.rpi.edu/web/project/MobileHealth
http://tw.rpi.edu/web/project/ReDrugS