

Extending Biology Models with Deep NLP over Scientific Articles

David McDonald, Scott Friedman, Amandalynne Paullada, Rusty Bobrow, Mark Burstein

SIFT, LLC

319 1st Ave North, Suite 400

Minneapolis, MN 55401

{dmcdonald, friedman, apaulada, burstein}@sift.net, rjbobrow@gmail.com

Abstract

This paper describes R3 (*Reading, Reasoning, and Reporting*), our system for deep language understanding and model management for the biomedical domain. Starting from a base BioPAX model, we learn extensions to it by reading biomedical research articles from PubMed Central. We describe the particular issues for text understanding in this domain and how we use pre- and post-analysis reasoning to bridge the differences in how knowledge is packaged in a text and in a biomedical database. We close with brief description of our first year results, where R3 was faster than all other reported systems, reading 1,000 articles in 15 minutes.

Introduction

Reading does not end with a parse or even with a semantic interpretation. When we read to inform ourselves, we use our current model of the world to guide our interpretation of the text, and then reconcile this interpretation with our original model. Our interpretation might corroborate, extend, or conflict with our world model and can cause us to revise or extend it. This concept of reading-with-a-model inspires our ongoing work on *Reading, Reasoning, and Reporting* (R3), as part of DARPA's "Big Mechanism" program (Cohen 2015). R3 reads articles in molecular biology to extend and revise its models of biological mechanisms.

Building a system that can read-with-a-model poses key research challenges for the general task and the specific domain:

- Heterogeneous ontologies in the domain model.
- The model continually changes due to new findings.
- Biology articles frequently discuss the *function* of events and entities, yet the most comprehensive biology models contain only *structural* data.
- The same word (e.g., "*Ras*") can refer to a protein, a gene, or a larger multi-protein complex, within a single article.
- A common reaction may be part of many other reactions in a domain model. Extending what we know about it requires us to accurately *localize* the correct instance within the model, i.e. to determine which instance of the reaction is the one referred to in the text.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

To address these challenges, R3 integrates deep semantic parsing, ontology mapping, and reasoning about structure, function, and mechanism-level causality. Deep parsing allows R3 to extract precise semantics and determine entity types from local lexical context. R3's ontology mapping allows it to query across heterogeneous ontologies within single rules to augment and localize the results of its semantic parse. R3's mechanism-level reasoning allows it to infer functional properties from structural descriptions and infer indirect causal relations to ground language to the model.

The R3 project is particularly concerned with inferring and exploiting the relationship between the text and the domain model, which have some practical inter-dependencies:

- *The model supports the interpretation.* The text rarely mentions sufficient properties or participants to uniquely identify an entity or a process; authors rely on the context and the reader's mental model of the domain to fill the gaps. The model provides context and targets for coreference and frame completion.
- *Model localization supports learning-by-reading.* The information in the text can only be used to improve/extend the model if it is properly localized within the model; otherwise, the model may be revised erroneously.
- *Model locales are interdependent.* Localizing text within the model can influence previous or subsequent localizations, since articles frequently describe sequentially or causally interdependent entities and events.

We begin by describing the problem of extracting and recognizing biological events and interactions from text, focusing on challenges for natural language understanding. We then describe the R3 approach to meeting these challenges and our empirical results that demonstrate R3's capabilities. We close with a discussion of the implications and future work for R3.

Machine Reading in the Biology Domain

Biomedical research articles are written to be read by other professional biologists who are presumed to have the requisite technical background. The brief mention of a well-known mechanism ("*Ras proteins*") is sufficient to evoke all of the details of the mechanism in the mind of the reader. This lets them effortlessly fill in information gaps that cannot be supplied by standard discourse techniques ("*activated*

upon GTP loading and deactivated upon hydrolysis of GTP to GDP” — loaded onto or hydrolyzed from what?). We need to have knowledge sources that let it do this too.

Like other authors, biologists are under pressure to keep their articles within length limits. This leads to compaction techniques such as describing events using nominalized verbs and packing information into them as prenominal modifiers, e.g. “*EGFR and ERBB3 tyrosine phosphorylation*,” “*mitogen-induced signal transduction*.” This changes the usual grammatical cues (such as one would use on newswire text) and requires knowledge-rich analysis techniques if parses are to be accurate.

A further property of biomedical text is that logically related information is usually distributed across multiple sentences. The example below is typical. The classification of the sites are given in the first sentence and their identity in the second. “*We observed two conserved putative MAPK phosphorylation sites in ASPP1 and ASPP2. The ASPP1 sites are at residues 671 and 746, and the ASPP2 sites are at residues 698 and 827.*” In R3 we have enhanced our discourse history to let us combine information from both sentences into a single, logically complete, representation.

Approach

Here we describe R3’s architecture and information flow. We use Figure 1 to guide our discussion, stepping through the information flow chronologically. We begin by describing the setup and operation of the domain model and the semantic parser, and then we discuss the post-parse reasoning mechanisms and operations on the domain model.

Bootstrapping the Domain Model and the Parser

Before reading articles, R3 initializes its parser with domain vocabulary and grammar and uses inference rules to optimize and index its domain model. R3 uses the UniProt knowledge base (UniProt Consortium 2008) as a source of protein synonyms to enhance protein recognition during parsing. It maps each protein synonym to a unique identifier to enable cross-indexing in various biological ontologies.

R3 imports OWL domain models specified in Biological Pathway Exchange (BioPAX) (Demir et al. 2010). BioPAX specifies structural information about biochemical reactions (e.g., bindings, phosphorylations, and other interactions), complexes, proteins, catalysis, and reaction regulation. R3 uses domain-specific inference rules to extend the BioPAX domain model with additional structure to explicitly represent causal relations, amino acids, functional information, and molecular categories (e.g., homo-dimer, heterodimer). We refer to these extensions as *enhanced BioPAX*. Much of the enhanced content is *implicitly* described in BioPAX (e.g., a homodimer is identifiable as a complex with two stoichiometry entities identifying the same protein), but R3 detects and explicitly represents this to facilitate its search and localization during reading.

Finally, R3 uses a graph grammar to segment and index the enhanced BioPAX model into different logical contexts. It uses the equivalent of regular expressions over its relational knowledge graph to describe how to traverse the

model and segment it into indexable parts, e.g., by starting with biochemical-reaction entities and traversing via left and right relations to their input and output molecules, respectively, and then descending recursively through sub-molecular structures via compound relations, etc.¹ This quickly segments and indexes the enhanced BioPAX model into smaller contexts so that R3 can quickly search the model to localize the information it reads.

Deep Semantic Parsing

The purpose of language analysis in R3 is to identify and represent the semantic content of biomedical texts to facilitate localization in the domain model and to provide a standard view of an article’s content for downstream reasoners (e.g. Danos et al. 2009). This entails normalizing all of the syntactic and lexical variation in how a relation is expressed to a single canonical form. Also, references to entities and relations must be aligned with articles’ document structure to facilitate search and context driven inferences.

To do this, R3 uses the SPARSER natural language analysis platform to read the texts. SPARSER is a rule-based, type-driven semantic parser. Rules succeed only if the types of the constituents to be composed satisfy the type constraints (value restrictions) specified by the rule. SPARSER is also model driven. As described in (McDonald 1996), writing a semantic grammar for it starts with a semantic model of the information to be analyzed along with a specification of all the ways each of the concepts can be realized in the language of the genre (e.g. biomedical research articles). A compiler takes the model and creates a semantic grammar from the realization specifications by drawing on a schematic standard English syntactic grammar. This ensures that everything SPARSER is able to understand (model) it can parse, and that every rule in the generated grammar has an interpretation.

R3 semantic interpretations are represented in a typed lambda calculus (McDonald 2000). The categories (predicates), are taken from an ontology (linguistically annotated domain model) whose upper structure is based on DOLCE (Gangemi et al. 2002) and Pustejovsky’s model of events (Pustejovsky 1991). There is a middle level with ontological models for location, time, people, measurement, change in amount, etc. This core is extended with a ontology of biomedical phenomena that is deliberately designed to be close to how these phenomena are described in articles in order to simplify the parsing process. Generalizations that are missed by parses that closely track the phrasing of a text can be compensated for by post-analysis reasoning.

Category instances – individuals – represent the entities, events, and relationships that are identified when a text is read. Individuals are unique: The parsing process guarantees that every individual with a particular set of values for its properties is represented by a single object (see McDonald 2000 and Maida and Shapiro 1982). This guarantee is managed by a description lattice that tracks the addition of properties (binding of variables). Every combination of

¹This graph grammar approach generalizes existing case-constructors (e.g., (Mostek, Forbus, and Mevarden 2000)).

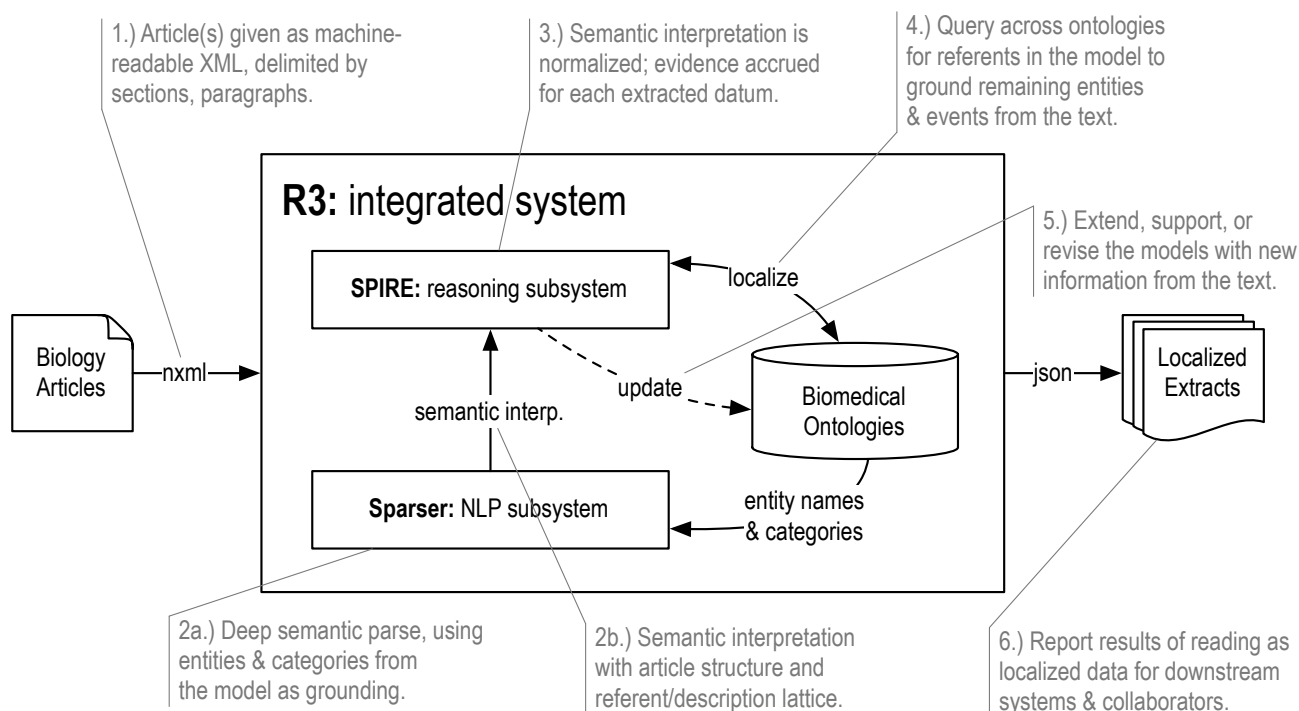


Figure 1: The R3 architecture, and the flow of information by which R3 reads articles, updates its mechanism models, and publishes extracted knowledge for human and machine collaborators.

property value and category instantiated is represented by a unique individual that is maintained and updated incrementally as a text is read.

Categories act as frames in a conventional knowledge representation, with a specialization lattice that permits the inheritance of realization options as well as variables (possible relations) and methods for type-specific reasoning. They are also where we state facts about normally expected properties. For example, phosphorylation events entail an active protein or other agent, the substrate protein that is phosphorylated, and the site (residue) where the phosphate is added. A residue is identified by its amino acid and its location on a particular protein. If we read about the sites of a phosphorylation and the requisite information is not supplied locally in the text, then we can assume that it is very likely to have been supplied elsewhere in the article, which motivates a search to identify it.

Our discourse component resolves pronominal and definite references using a structured history of entity and event mentions. This same facility organizes searches to expand partial descriptions of entities to full ones (frame completion) and in general to link individuals as they appear in different parts of an article. Consider this text. It compares what happens when a particular drug is or is not used:

“In untreated cells, EGFR is phosphorylated at T669 by MEK/ERK, which inhibits activation of EGFR and ERBB3. In the presence of AZD6244, ERK is inhibited and T669 phosphorylation is blocked, increasing EGFR and ERBB3

tyrosine phosphorylation and up-regulating downstream signaling.”

There are two mentions of the phosphorylation of residue T669 in this text, one in each sentence. The mention in the second sentence (“T669 phosphorylation”) is marked by the sentence post-processor as being incomplete because it does not specify the agent or the substrate. This combination of event-type and site is a unique individual stored in the description lattice. The discourse history records that this individual was also mentioned in the first sentence. This is enough to license R3 to trace up the structure on the first mention to identify the other properties it has, and to copy over any non-conflicting properties of the first to the second.²

Localizing Against the Model

R3 next localizes the extracted information within its domain model. Since the parser does not directly produce BioPAX, R3 maps the extracted information into the enhanced BioPAX ontology to support localization. This involves translating across relational/category vocabularies as well as generating new symbols to account for changes in event/entity granularity across ontologies.

After mapping the extracted information into enhanced BioPAX, R3 uses it as a probe to search the entire do-

²The two eventualities differ in their existential status. The tense in the first sentence indicates that the phosphorylation occurs. In the second we are told that it is blocked.

main model. R3 uses a two-stage similarity-based retrieval algorithm (Forbus, Gentner, and Law 1995), starting with a quick feature vector comparison between the probe and each model context, and culminating with a graph-matching algorithm, similar to structure-mapping (Friedman 2015; Falkenhainer, Forbus, and Gentner 1989) but with a strong preference for literal identity over just relational similarity.

R3 thereby identifies and ranks portions of the domain model according to their similarity to the extracted knowledge. This graph-matching approach has the following benefits:

- **Multiple matches:** if the article *omits* something from the model, which is almost always the case, R3 will retrieve multiple relevant candidates for additional consideration.
- **Partial matches:** if the article mentions something *not* in the model (e.g., a novel regulatory process) the description of the surrounding context (e.g., molecules and biochemical reactions) will still be present in the probe to help R3 retrieve relevant portions of the model to extend.
- **Inference:** R3's graph-matching process computes *candidate inferences* for transferring unmatched entities and relations from the article into the domain model, whether the new knowledge conflicts with the model or extends it. In previous work, we have shown that these inferences can be practically used to revise beliefs and models (Burstein 1988; Friedman, Barbella, and Forbus 2012).

Semantic similarity is not sufficient to uniquely identify referents from the text. Consider the sentence "*SOS and Grb2 promote the formation of GTP-bound p21 Ras.*" Without more information, this will perfectly match at least 13 distinct biochemical reaction entries in R3's BioPAX model.

Distinguishing which of these perfect matches the article refers to—and it could be more than one—R3 must use the context of the surrounding article text. We are implementing a measure of *causal relevance*, so R3 can use previous, high-confidence localization operations to rank these candidates based on their proximity in the causal model. This assumes that biology articles describe causally-related events and entities rather than unrelated events and entities, which holds true in our experience.

At present, R3 updates the model by extending the enhanced BioPAX model with new information (dotted arrow, Figure 1). Important near-term future work on R3 will enable it to automatically identify possible conflicts, pose resolutions to these conflicts, and retain provenance in order to allow intervention by human experts.

Evaluation

We evaluated R3's semantic parser and its ability to extract information, filter irrelevant information (i.e., entities or events not in the domain model) and merge duplicate information against 1,000 biology articles from PubMed Central provided by the Big Mechanism Program.

We configured R3 to extract information about phosphorylation reactions, ubiquitination reactions, positive and negative regulation of processes, and increases or decreases

in molecule concentrations. Other information—including binding events, indirect causal relations, translocation events, transcription events, and more—were parsed but not analyzed with respect to the domain model. Additionally, R3 used epistemic filtering to ignore historical, hypothetical, or negated statements, in order to focus on positive information.

R3 read all 1,000 articles in 15 minutes. In total, it extracted 15,876 semantic descriptions of the targeted data, across all sections of all papers. This includes entities and events that were unrelated to the model, as well as duplicate data, since multiple sentences often refer to the same event.

R3 discarded 619 data that were only mentioned in the introduction or methods sections, since it is designed to focus on the contributions of articles, and not the exposition or methodology. It then analyzed each extracted datum for model relevance, e.g., whether the proteins of a reaction are described in the domain model. R3 filtered out 8,864 irrelevant data, leaving 6,384. Finally, R3 merged these entities and events—and the parsed text that served as evidence—into 2,351 data pertaining to the domain model.

Conclusion & Future Work

This paper outlined how R3 reads scientific articles to improve its scientific models. Our development and evaluation of R3 to date has focused mostly on the semantic parsing and model localization work. R3 can presently read articles, extract knowledge, localize extracted knowledge within the model, and determine which extracted data support the model and which extend the model, but R3 does not yet revise the model based on these extensions.

Going forward, we will increase R3's competence in reading and model manipulation across the board. In particular we aim to construct a set of 'mini' mechanism models to provide the machine equivalent of the knowledge that authors take for granted. To get this knowledge we are starting to machine-read the comments that curators make about database reactions. We also expect to enlist biologists to write simple mechanism descriptions for R3 to read. This sort of knowledge-bootstrapping is challenging, but we feel it is the only way we will be able to handle the breadth of new research in molecular biology and contribute to advances in biomedicine.

Acknowledgments

This work was supported in part by the DARPA "Big Mechanism" program, BAA 14-14, under ARO contract W911NF-14-C-0109. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the position or policy of the Government.

References

- Burstein, M. H. 1988. Combining analogies in mental models. In *Analogical Reasoning*. Springer. 179–203.
- Cohen, P. R. 2015. DARPA's big mechanism program. *Physical Biology* 12.

- Danos, V.; Feret, J.; Fontana, W.; Harmer, R.; and Krivine, J. 2009. Rule-based modelling and model perturbation. In *Transactions on Computational Systems Biology XI*. Springer. 116–137.
- Demir, E.; Cary, M. P.; Paley, S.; Fukuda, K.; Lemer, C.; Vastrik, I.; Wu, G.; D’Eustachio, P.; Schaefer, C.; Luciano, J.; et al. 2010. The biopax community standard for pathway data sharing. *Nature biotechnology* 28(9):935–942.
- Falkenhainer, B.; Forbus, K. D.; and Gentner, D. 1989. The structure-mapping engine: Algorithm and examples. *Artificial Intelligence* 41(1):1 – 63.
- Forbus, K. D.; Gentner, D.; and Law, K. 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19(2):141–205.
- Friedman, S. E.; Barbella, D. M.; and Forbus, K. D. 2012. Revising domain knowledge with cross-domain analogy. *Advances in Cognitive Systems* 2:13–24.
- Friedman, S. E. 2015. Exploiting graph structure to summarize and compress relational knowledge. *Proceedings of the 28th International Workshop on Qualitative Reasoning*.
- Gangemi, A.; Guarino, N.; Masolo, C.; Oltramari, A.; and Schneider, L. 2002. Sweetening ontologies with DOLCE. In *Knowledge engineering and knowledge management: Ontologies and the semantic Web*. Springer. 166–181.
- Maida, A., and Shapiro, S. 1982. Intensional concepts in propositional semantic networks. *Cognitive Science* 6:291–330.
- McDonald, D. D. 1996. The interplay of syntactic and semantic node labels in partial parsing. In Bunt, H., and Tomita, M., eds., *Recent Advances in Parsing Technology*. Kluwer Academic Publishers. 295323.
- McDonald, D. D. 2000. Issues in the representation of real texts: The design of Krisp. In Iwanska, L. M., and Shapiro, S. C., eds., *Natural Language Processing and Knowledge Representation*. MIT Press. 77–110.
- Mostek, T.; Forbus, K. D.; and Meverden, C. 2000. Dynamic case creation and expansion for analogical reasoning. In *AAAI/IAAI*, 323–329.
- Pustejovsky, J. 1991. The syntax of event structure. *Cognition* 1(41):47–81.
- UniProt Consortium. 2008. The universal protein resource (uniprot). *Nucleic acids research* 36(suppl 1):D190–D195.