Cycles of Scientific Investigation in Discourse

Machine Reading Methods for the Primary Research Contributions of a Paper

Gully A. Burns
ISI, USC
Marina del Rey, CA, USA
gully@usc.edu

Anita de Waard
Elsevier Research Data Services
Jericho, VT, USA
a.dewaard@elsevier.com

Pradeep Dasigi
LTI, CMU
Pittsburgh, PA, USA
pdasigi@cs.cmu.edu

Eduard H. Hovy
LTI, CMU
Pittsburgh, USA
hovy@cmu.edu

***Abstract*— We describe a novel approach to machine reading of the primary scientific literature. We treat a description of an experiment as a discourse, viewing a scientific corpus not merely into a collection of documents, but also an extended conversation formed by the collective set of experiments, their introductions and interpretations. This paper introduces this approach as a methodology called 'Cycles of Scientific Investigation in Discourse' (CoSID). In CoSID, we capture the central conceptual structure of a paper as a series of nested reasoning loops, composed of passages in results sections, which describe individual research findings. We ground our work with a number of worked examples based on data from the MINTACT and Pathway Logic databases, and illustrate the idea in the context of machine-enable biocuration[1].**

Keywords—interpretive framework for experiments, experiment description as discourse, computational language technology

I. Theoretical Background

All experiments consist of a series of actions performed upon entities, conducted for a reason, ending with a measurement/evaluation of something and an interpretation as conclusion. But people do not conduct experiments in a vacuum. Experiments are formulated to explore possibilities within a larger encapsulating theory, and their conclusions are intended to flesh out the unknown parts of the theory. They can therefore be viewed as 'knowledge turns' in an ongoing discourse, with internal linkage among corresponding portions (specific goals, hypotheses, conclusions, etc.).

Experiments, by their nature, are specific: actions situated in time and space, performed with physical objects. Theories, in contrast, are by their nature general, intended to apply beyond the particular time and place of the experiment. They employ abstractions that any particular experiment has to instantiate as its artifacts and activities. Since theories are 'conceptual' while experiments are 'practical' in nature, it may be very difficult for an experiment to serve as an absolute proof for any theory for all time and space.

CoSID (Cycles of Scientific Investigation in Discourse) is a model of experimentational text that takes into account these two points of view. If we postulate that that scientific investigation proceeds in cycles of increasing theoretical specificity (each round of experiments serving to inform the next round of conceptual expansion), the CoSID model provides a formalization that abstracts from the text of scientific papers to a set of representations that support cross-paper tracking, comparison of ideas, hypothesis evolution, etc. This facilitates the understanding of how experimentally-founded knowledge is created and developed over time and space by a disjointed scholarly community, through processes of reading, writing, and experimentation.

To capture how experiments are presented in technical publications we create in CoSID three layers of representation, each being a frame with associated properties:

1. **Context** — the conceptual framework about some phenomenon. In principle this exists 'outside' any particular paper, but for any paper, it provides the framework for all experiments within it (and also forms a localized context for experiments from a single section of a paper). We model this with a computational frame structure that includes slots for hypotheses, pointers to experiments, a description of the overall interrelation of experiments and interpretation, etc.

2. **Experiment** — a series of physically instantiated activities governed by a goal and hypothesis, resulting in observations and measurements. Generally a technical paper containes many experiments (each possibly only briefly described). Each one explores some specific combination of parameter values, and is modeled by a frame whose slots provide the goal, method, observed results, specific experimental implications, etc.

3. **Interpretation** — the interpretations drawn from one or more experiments, leading back to the overall interpretation in the Context (above). Each experiment's local hypothesis makes up a part of the global hypothesis of the Context.

We represent a CoSID frame as a nested structure where a single Context associates with multiple Experiments and is concluded with a single Interpretation. Each CoSID frame is derived from a passage in the results section that points to subfigures that each report individual experiments. Figure 1 shows the application of CoSID to a sample article (pmid: 10533201) where the discourse structure of a single frame (Fig1AB) is explored. This frame consists of 12 clauses moving from facts to methods, results, and interpretation, to inform the frame structure as described above.

II. Corpora and Data

Our overall goal is to produce automatically for a given scientific paper a set of instantiated CoSID frames, all properly connected, that completely and accurately reflect its contents. To this end, we have to perform multiple quite distinct tasks, including determining the overall goals, backgroumd, and hypotheses of the paper, identifying where individual

[1] This work was funded by DARPA Big Mechanism program under

experiment boundaries lie, understanding each of them individually, connecting everything together, and then creating the appropriate interlinked frame structures.

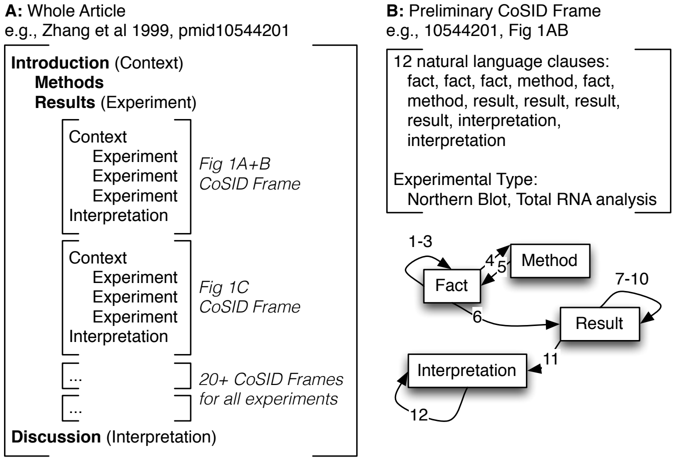

Figure 1: Applying CoSID frames to a research article. A: Overall structure of frames within the textual narrative, B: Discourse structure within a frame showing transition between discourse types.

This work is performed in the RUBICON project, funded by DARPA's Big Mechanisms program, that is extracting relevant facts from a vast collection of papers about Ras cancers and formulating them to support theoretical model builders, automated reasoners, and actual experimenters [1]. Our contribution is to provide rich contexts in which individual atomic statements about biological entities, extracted by others, can be properly interpreted (for example, as hypothetical or as actual, or as a local interpretation drawn from an experiment, or one drawn from some other work reported).

We focus on the text associated with the subfigure (i.e., Fig. 1A, 3C, etc.) and develop classifiers for the type of experiments performed. To test our work, we use two manually curated models of the data: The Pathway Logic group at SRI International contain approximately 2,000 papers of which 76 are open-access. Each data record is assigned one of 33 separate 'assay types'[2] (such as 'coprecipitation', 'phosphoryation', etc). Similarly, the MINTACT database provides hand-curated records of 37,268 experiments from 14,009 papers, of which 1,063 are available as open access papers [2].

III. Work to Date

Our first step is to delimit each experiment. We accomplish this by processing the caption of each Results section figure. Accuracy within captions is essentially perfect, given helpful phrases like "Figure 2(a) depicts…". Using this, we search within the Results section text to find a reference to the corresponding portion of the figure, as in "As shown in Fig 2(a),…". This forms the anchor for a span of text that, we assume, provides details about a single experiment. We trained a Conditional Random Field (CRF) model to assign types to these experiments (from either the PSI-MI2.5[3] or the Pathway Logic typology). Separately, we parse the text into discourse segments, each roughly a clause, and identify for each one a Discourse Segment type, along the lines of [3]. These types include the labels 'fact', 'hypothesis', 'problem', 'goal', 'method', 'result', and 'interpretation'. As a third component, we are working to identify the theoretical model that underlies each paper, which will form part of the Context frame.

IV. Early Results

To date, we have implemented several modules, including: (A) A caption splitter that uses rules to identify individiual experiments inside captions. Performance is >95%. (B) An experiment delimiter that uses rules to delimit the extent of each experiment description in the Results section of the paper. (C) An experiment type tagger. We experimented with different numbers of types, sometimes condensing the less-frequent ones together (F1-score: 71%). (D) A discourse segment type tagger, a trained CRF model to assign a discourse segment tag to each clause (F1-score: 66%).

V. Next Steps

This work is ongoing. After completing the missing components we plan to make available a collection of Ras cancer papers with associated CoSID frames, which we believe will be useful within the FRIES consortium in the Big Mechanism program. FRIES includes groups building systems that extract atomic information about entities and relations from papers about Ras cancer research, individuals creating models of Ras cancer and associated experiments, and groups building automated modeling and reasoning systems.

Some uses for our work include: downweighting the certainty score for assertions that have been tagged as *hypotheses*, compared to *facts*; downweighting the assertions from high-level conclusions, as compared to level-level direct experimental findings, since the former may suffer from misconstrual; allowing models to cross-link experiments from different papers when their Experiment frames are similar enough (i.e., they apply the same experimental techniques in the same settings to the same materials); and more.

We welcome suggestions for additional uses and extensions of the CoSID model.

References

[1] Cohen, P.R. DARPA's Big Mechanism program. *Phys Biol* **12,** 045008 (2015)

[2] Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358–363 (2014).

[3] de Waard, A. and Pander Maat, H.L.W., Verb form indicates discourse segment type in biological research papers: Experimental evidence *Journal of English for Academic Purposes*, 11 (4), (pp. 357-366), doi:10.1016/j.jeap.2012.06.00

[2] http://pl.csl.sri.com/CurationNotebook/pages/Assays.html

[3] http://www.psidev.info/node/60