The CISP Annotation Schema Uncovers Hypotheses and Explanations in Full-Text Scientific Journal Articles

Elizabeth White, K. Bretonnel Cohen, and Lawrence Hunter
Department of Pharmacology, Computational Bioscience Program,
University of Colorado School of Medicine, Aurora, Colorado, USA
elizabeth.white@ucdenver.edu,
kevin.cohen@gmail.com,
larry.hunter@ucdenver.edu

Abstract

Increasingly, as full-text scientific papers are becoming available, scientific queries have shifted from looking for facts to looking for arguments. Researchers want to know when their colleagues are proposing theories, outlining evidentiary relations, or explaining discrepancies. We show here that sentence-level annotation with the CISP schema adapts well to a corpus of biomedical articles, and we present preliminary results arguing that the CISP schema is uniquely suited to recovering common types of scientific arguments about hypotheses, explanations, and evidence.

1 Introduction

In the scientific domain, the deluge of full-text publications is driving researchers to find better techniques for extracting or summarizing the main claims and findings in a paper. Many researchers have noted that the sentences of a paper play a small set of different rhetorical roles (Teufel and Moens, 1999; Blais et al., 2007; Agarwal and Yu, 2009). We are investigating the rhetorical roles of sentences in the CRAFT corpus, a set of 97 full-text papers that we have annotated using the CISP schema. Hand alignment of the resulting annotations suggests that patterns in these CISP-annotated sentences correspond to common argumentative gambits in scientific writing.

2 Methods

The CRAFT corpus is a set of 97 full-text papers describing the function of genes in the Mouse Genome Informatics database (Blake et al., 2011). These documents have already been annotated with syntactic information (parse trees and part-of-speech tags), linguistic phenomena (coreference), and semantic entities (genes, chemicals, cell lines, biological functions and molecular processes), making the corpus a rich resource for extracting or inferring information from full scientific papers.

The CISP schema (Soldatova and Liakata, 2007; Liakata et al., 2009) contains 11 categories, and several of the categories describe the intentions of the authors, making it well suited for markup of argumentation. We chose to narrow these down to 9 categories (excluding Model and Object) during annotation training; our guidelines are shown in Figure 1. We expect this schema to describe the pragmatics in the text well, while still offering the potential for high interannotator agreement due to a manageable number of categories. The process of marking the sentences in the CRAFT corpus according to the CISP guidelines took one annotator about four months.

3 Results and Discussion

Six of the 97 CRAFT papers do not follow the standard IMRaD paper structure (one was a review article, and five combined Results and Discussion); these documents were eliminated from this analysis. Annotation of the 91 remaining CRAFT papers resulted in 20676 sentences. The distribution of the annotated classes is shown in Table 1.

Our use of the CISP schema exposes an approach for recovering two types of explanatory arguments. The first sets the context with a sequence of Back-

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 134–135,
Portland, Oregon, USA, June 23-24, 2011.

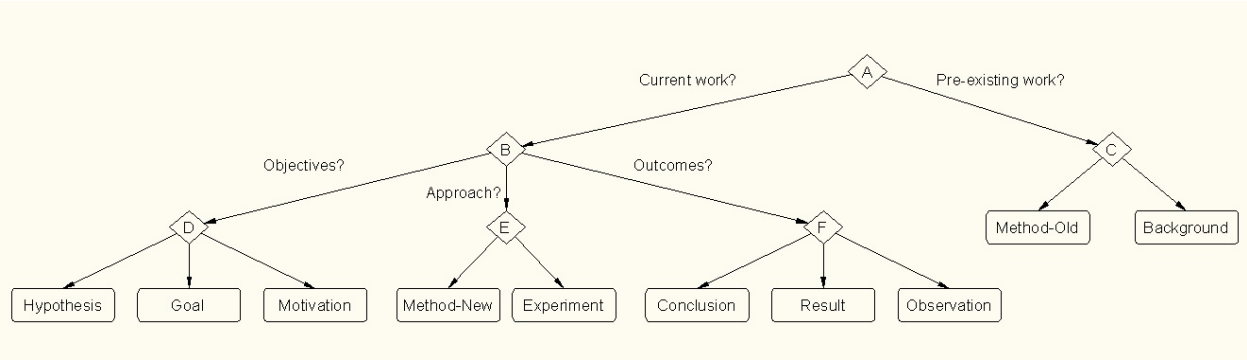

Figure 1: Flow chart for CISP annotation of the CRAFT corpus.

CISP Type	Count	Percentage
Hypothesis	1050	5.08
Goal	992	4.80
Motivation	928	4.49
Background	2838	13.73
Method	637	3.08
Experiment	5270	25.49
Result	5471	26.46
Observation	1168	5.65
Conclusion	2322	11.23
Total	20676	100.0

Table 1: Distribution of CISP sentence types annotated in 91 CRAFT articles.

ground sentences, followed by a Hypothesis, Motivation, or Goal; this echoes a motif found by Swales (1990) and Teufel and Moens (1999). We also find another pattern that consists of a combination of Results and Observations, either preceded or followed by a Conclusion; Teufel and Moens (1999) also find exemplars of this maneuver, and note that it parallels Swales' notion of occupying a niche in the research world. Hand alignment of CISP annotations in Introduction and Result sections suggests that a finite state machine may be capable of modeling the transitions between CISP sentence types in these arguments, and machine learning approaches to represent these and other patterns with hidden Markov models or conditional random fields are underway.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results, and Discussion. *Bioinformatics*, 25(23): 3174–3180.

Antoine Blais, Iana Atanassova, Jean-Pierre Desclés, Mimi Zhang, and Leila Zighem. 2007. Discourse automatic annotation of texts: an application to summarization. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, May 7-9, 2007, Key West, Florida, USA, 350–355. AAAI Press.

Judith A. Blake, Carol J. Bult, James A. Kadin, Joel E. Richardson, Janan T. Eppig, and the Mouse Genome Database Group 2011. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, 39(Suppl. 1): D842–D848.

Maria Liakata, Claire Q, and Larisa N. Soldatova. Semantic Annotation of Papers: Interface & Enrichment Tool (SAPIENT). 2009. In *Proceedings of BioNLP 2009*, Boulder, Colorado,193–200.

Larisa Soldatova and Maria Liakata. 2007. An ontology methodology and CISP - the proposed Core Information about Scientific Papers. JISC intermediate project report.

John M. Swales. 1990. *Genre Analysis: English in academic and research settings*, 137–166. Cambridge University Press, Cambridge.

Simone Teufel and Marc Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In *Advances in Automatic Text Summarization*, I. Mani and D. Maybury, eds. MIT Press, Cambridge, MA.