Ontological Knowledge for Rhetorical Move Analysis

Mohammed Alliheedi[1,3], Robert E. Mercer[2,3] and Sandor Haas-Neil[2]

Al Baha Unversity[1]
The University of Western Ontario[2]
University of Waterloo[3]
mallihee@uwaterloo.ca
mercer@csd.uwo.ca
shaasnei@uwo.ca

Abstract. Scholarly writing in the experimental biomedical sciences follows the IMRaD (Introduction, Methods, Results, and Discussion) structure. Many Biomedical Natural Language Processing tasks take advantage of this structure. Recently, a new challenging information extraction task has been introduced as a means of obtaining these types of detailed information: identifying the argumentation structure in biomedical articles. *Argumentation mining* can be used to validate scientific claims and experimental methodology, and to plot deeper chains of scientific reasoning. One subtask in identifying the argumentation structure is the identification of *rhetorical moves*, text segments that are rhetorical and perform specific communicative goals, in the Methods section. Based on a descriptive taxonomy of rhetorical moves structured around IMRaD, the foundational linguistic knowledge needed for a computationally feasible model of the rhetorical moves is described: *semantic roles.* One goal is to provide FrameNet and VerbNet-like ontologies for the specialized domain of biochemistry. Using the observation that the structure of scholarly writing in the laboratory-based experimental sciences closely follows the laboratory procedures, we focus on the procedural verbs in the Methods section. Occasionally, the text does not contain fillers for all of the semantic role slots that are needed to perform an adequate analysis of a verb. To overcome this problem, an ontology of experimental procedures can be interrogated to provide a most likely candidate for the missing semantic role(s).

1 Introduction

Scientists must routinely review the scholarly literature in their fields to keep abreast of current advances and to retrieve information relevant to their research. However, the volume of online scientific literature is immense, and rapidly increasing. In the biomedical field, the National Centre for Biotechnology Information (NCBI) developed a literature search engine, PubMed[1], to access various

[1] http://www.ncbi.nlm.nih.gov/pubmed

databases such as MEDLINE (journal citations and abstracts for biomedical literature), full-text life science e-journals, and online books. In 2010 PubMed repositories consisted of more than 20 million citations for biomedical literature [18]. By 2015 the number of citations had increased to more than 25 million[2]. As a consequence, it has become extremely challenging for biomedical scientists to keep current with information in their fields. This challenge has attracted Natural Language Processing (NLP) researchers to develop resources and automated tools for performing various tasks in Information Extraction (IE) and Text Mining (TM) using online corpora of biomedical articles, and thus enable biomedical researchers to better manage and exploit this volume of data [13]. These research activities have led to the development of a new field, Biomedical Natural Language Processing (BioNLP), a collaboration between the biomedical and computational linguistics/artificial intelligence communities [12].

The types of tasks currently handled by BioNLP systems have generally been aimed at extracting very specific and limited information, for example, protein and gene names and relations [6], and so have been able to rely on relatively simple forms of information extraction. BioNLP has adapted various standard information extraction techniques, including both rule-based (e.g., shallow parsing, syntactic pattern-matching) and Machine Learning (e.g., Support Vector Machines, k-nearest neighbour classification method), to address several text-mining tasks, including extracting: protein-protein interactions (PPI) [16], drug-drug interactions (DDI) [23], gene relationships [14], and protein-residue associations [20].

Although these approaches fulfil some information needs, information extraction systems based on these can only recognize and extract minimal and specific information from biomedical texts. But other, more in-depth and comprehensive, information contained in biomedical texts would be highly valuable to scientists because this type of information can enable validating scientific claims, tracing current research directions in their field, reproducing scientific procedures and so forth. Recently, a new and more challenging information extraction task has been introduced as a means of obtaining these types of detailed information: identifying the argumentation structure in biomedical articles (e.g., [10] and [11]). *Argumentation mining* can be used to validate scientific claims and experimental methodology, and to plot deeper chains of scientific reasoning. Unlike earlier simpler forms of information extraction, here the goal is to identify the structure of argumentative components within an entire text—for example, premises, evidence, conclusions—as well as the relationships between components.

To achieve this goal the text needs to be analyzed. Our approach to this analysis is based on a working hypothesis: We hypothesize that recognizing and detecting rhetorical moves would provide important information to our argumentation analysis framework, and that the Method sections in biochemistry articles contain moves which can be correlated with the author's experimental procedures. These moves can be used to determine salient information about the elements of the article's argumentative structure (e.g., premises) and can

[2] http://www.ncbi.nlm.nih.gov/books/NBK3827/

contribute to the overall understanding of the author's scientific claims. A key aspect of our hypothesis is that development of a frame-based knowledge representation can be based on the semantics of the verbs associated with these procedures. This representation can provide detailed knowledge for understanding these rhetorical moves, which will in turn facilitate analysis of argumentation structure. In other words, we propose that a *procedurally rhetorical verb-centric frame semantics* can be used to obtain a sufficiently deep analysis of sentence meaning .

While this approach seems straightforward enough, the writing style of biochemistry articles requires the reader to have knowledge about biochemistry and biochemistry laboratory techniques and practices. This paper first gives the semantic roles that can be used in the semantics of each verb. Then an example of how an ontology containing knowledge about biochemistry laboratory techniques and practices can be used to fill the semantic roles of verbs which cannot be filled by information in the text.

2 Related Work

Swales [24] proposed the Create-A-Research-Space (CARS) model that uses intuition about the argumentative structure of scientific research articles. Swales defined rhetorical moves as text segments that convey communicative goals. He reviewed the Introduction section in 48 articles from social and natural science and found common rhetorical structures among most of these articles. Swales identified three moves in these articles: establishing a research territory, establishing a niche, and occupying the niche. However, despite the widespread influence of the CARS model, some researchers observed two problems: (i) the inconsistent assignment of rhetorical moves to text segments because the identification of the rhetorical moves relies on overall text comprehension, and (ii) a lack of empirical validation of moves in linguistic terms [15].

To overcome these problems, Kanoksilapatham [15] advanced Swales' approach to move analysis by developing a framework that combines his original CARS model with the use of Biber's multidimensional analysis [2] to enrich the model with additional information about linguistic characteristics. Biber's multidimensional analysis [2] is concerned with variation in the speaking and writing of English. Multidimensional analysis can be used to identify differences in linguistic characteristics between various text types at different levels of document structure (e.g., genre, internal section level). Although Kanoksilapatham provides an extension to the Swales's move analysis study, and attempted validation of these moves in biochemistry articles, she only provides a descriptive analysis about rhetorical moves without defining an explicit method for analyzing and recognizing these moves in texts.

Liakata et al. [17] developed an annotation scheme called Core Scientific Concepts (CoreSC) to classify sentences into scientific categories (e.g., related to author's other work). The CoreSC scheme consists of three layers: the first includes several categories to classify sentences; the second layer is concerned with properties of these categories; and the third layer creates a link to related instances of the same category. The authors use Machine Learning classifiers (i.e., Conditional Random Fields and Support Vector Machines) to automatically classify sentences into the CoreSC categorizes. The data set consisted of 265 biochemistry and chemistry articles. The authors were only able to achieve an accuracy around 50% in categorizing sentences in the appropriate CoreSC scientific categories which is inadequate for such a task.

Green [10] proposed a plan for creating an annotated corpus of biomedical genetics research articles. Green emphasized that this corpus would be beneficial to the argumentation mining community since it would provide a fine-grained annotation of argumentative components. Also since there are as yet few annotated corpora available, such a corpus would enrich research in the field of Computational Argumentation in general. The author stated that this corpus will be publicly available for further investigation by different research groups in various tasks of argumentation mining.

Green [11] specified a set of argumentation schemes for scientific claims in genetics research articles. The author used a corpus of unannotated genetics research articles, and identified the components (e.g., premises, conclusions) of an argument as well as its type of scheme. Based on the analyses of various genetics research articles, the author specified 10 argumentation schemes that are semantically different. These schemes were new and had not previously been proposed. Furthermore, the specification of argumentation schemes was used to create annotation guidelines. Then, these guidelines were evaluated in a pilot study based on participants' ability to recognize these schemes by reading the guidelines. Overall, the author's ultimate goal for this initial study was to develop annotation guidelines for creating corpora for argumentation mining research. However, based on the pilot study, the results showed a variation in performance since there were two groups of participants (i.e., undergraduate students and researchers). The students performed poorly in recognizing argumentation schemes while the researchers were able to identify these schemes correctly in most cases.

3 Our Proposed Approach: Argumentative Moves Mirror Scientific Experimental Procedures

Our intention is to develop a formal knowledge representation based on procedural verbs as a method for argumentation analysis. We introduced the notion of Swale's CARS model [24] in Section 2. We hypothesize that recognizing and detecting argumentative moves would provide additional information to our framework of argumentation analysis. We also hypothesize that the Method sections in biochemistry articles contain moves which can be correlated with the author's experimental procedures. These moves can be used to determine salient information about the elements of the article's argumentative structure (e.g., premises) and can contribute to the overall understanding of the author's scientific claims. A key aspect of our hypothesis is that development of a frame-based knowledge representation can be based on the semantics of the verbs associated with these procedures. This representation can provide detailed knowledge for understanding these argumentative moves, which will in turn facilitate analysis of argumentation structure. In other words, we propose that a *procedurally rhetorical verb-centric frame semantics* can be used to obtain a deeper analysis of sentence meaning than is currently the case with simple methods of Information Extraction (e.g., shallow syntactic pattern) and in a computationally feasible manner.

Scientific argument[3] is defined as a process that scientists follow by using certain procedures to obtain empirical data which will either support or defeat their claims, hence leading to the intended conclusion. The strength of a scientific argument depends on its reproducibility and consistency. For a scientific argument to be strong, a scientist should identify and explain all the procedures in their experiment, i.e., reproducibility, so that another researcher who follows the same procedures will reach the same conclusion, i.e., consistency. Thus, for a well-constructed scientific article, a scientist should expect the same conclusion if she follows the same procedures in the same sequence as described in the Method section.

Scientific writing in the biochemistry domain has certain characteristics that made it ideal for our purposes. In this domain, experimental procedures describe the sequence of actions the biochemist performs to carry out an experiment to derive scientific conclusions, to demonstrate science experiments as can be seen in the experimental manuals (e.g., Boyer [3] and Sambrook and Russell [21]). Verbs play an essential role as indicators of these experimental procedures. These procedures can be viewed as corresponding to the elements of the scientific argumentation structure. For example, when examining a biological substance (e.g., a certain type of bacteria) in order to prove a hypothesis (e.g., this bacteria is correlated with a certain disease) the biochemist would perform a sequence of certain procedures to arrive at a conclusion. Essentially, biochemists create an argumentation framework through the scientific methodology they follow—how they perform their experiments is how they argue. We can observe that this genre— biochemistry articles—is procedure-oriented since the scientific procedures that are described are parallel to the scientific argumentation in the text. For example:

Example 1. “Beads with bound proteins were washed six times (for 10 min under rotation at 4 C) with pulldown buffer and proteins harvested in SDS-sample buffer, separated by SDS-PAGE, and analyzed by autoradiography.” [7].

In this example, the verbs “washed”, “harvested”, “separated”, and “analyzed” are used to illustrate the procedure steps in sequential order. Such an experiment can be reproduced if one follows these steps.

Fillmore [8] introduced the notion of frame semantics as a theory of meaning. A *semantic frame* is defined as “any coherent individuatable perception,

[3] http://www.ces.fau.edu/nasa/introduction/scientific-inquiry/why-do-scientists-argue-and-challenge-each-others-results.php

Move type	Definition
Description-of-method	Concerned with sentences that describe experimental events.
Appeal-to-authority	Concerned with sentences that discuss the use of well-established methods.
Background information	Concerned with all background information for the experimental events such as "method justification, comment, or observation, exclusion of data, approval of use of human tissue" as defined by Kanoksilapatham (2003).
Source-of-materials	Concerned with the use of certain biological materials in the experimental events.

Table 1. Rhetorical Moves in the Method Section of Biochemistry Articles (from [15])

memory, experience, action or object" by Fillmore [9]. In other words, coherently structured concepts that are related to each other represent a complete knowledge of world events or experiences. For example, to understand the word "buy", one would access the knowledge contained in the commercial transaction frame which includes words such as the person who buys the goods (buyer), the goods that are being sold (goods), the person who sells the goods (seller), and the currency that the buyer and seller agree on (money).

Following Fillmore's theory of frame semantics, FrameNet [1] was developed to create an online lexical resource for English. This framework includes more than 170,000 manually annotated sentences and 10,000 words. The computational linguistic community has been attracted to the concept of the frame semantics and developed computational resources using this concept, such as VerbNet [22], an on-line verb lexicon for English and PropBank [19], an annotated corpus with basic semantic propositions.

Following the notion of frame semantics, we propose to build a knowledge representation framework to analyze verbs in a procedure-oriented genre. Our concept of procedurally rhetorical verb-centric frame semantics is intended to address this gap by developing a computationally feasible knowledge representation that will enable argumentation analysis. The knowledge contained in the frame semantics will facilitate the extraction of elements of arguments, i.e., argumentation mining. To reiterate, our hypothesis is that procedurally rhetorical verb-centric frame semantics can provide a knowledge representation framework for analyzing and representing the meanings of the verbs used in biochemistry articles. In turn, these frames will facilitate the identification of argumentation structure in the discourse describing experimental procedures.

3.1 Semantic Roles

As described earlier our experimental event scheme was inspired by the annotation scheme for bio-events [25]. We based our experimental event scheme for verb arguments on the inventory of semantic roles in VerbNet [22] and modified and added new semantic roles to define our scheme. Our experimental event scheme

Semantic role	Definition
Agent	"Generally a human or an animate subject. Used mostly as a volitional agent, but also used in VerbNet for internally controlled subjects such as forces and machines"[4].
Patient	"used for participants that are undergoing a process or that have been affected in some way"[5].
Predicate	A word that initiates the frame. It could be a verb such as *compare*, or a nominalized verb such as *transcription* or *activation.*
Theme	"used for participants in a location or undergoing a change of location"[6].
Goal	Identifies a thing toward which an action is directed or place to which something moves [7].
Factitive	"An referent that results from the action or state identified by a verb" [8].
Location	The physical place where the experiments took place.
Protocol-Detail:Time	Identifies the time or a duration of an experimental process.
Protocol-Detail:Temperature	Identifies the temperature of an experimental process.
Protocol-Detail:Condition	Identifies the condition of how an experimental process being carried out (e.g., under rotation).
Protocol-Detail:Repetition	Identifies the number of times that an experimental process being repeated.
Protocol-Detail:Buffer	Identifies the buffer that was used in an experimental process.
Protocol-Detail:Cofactor	Identifies the cofactor that was used in an experimental process.
Instrument:Change	Describes an object or protocol that can change another object(s). This role corresponds closely with the VerbNet project[9] instrument semantic role which describes something "used to describe objects (or forces) that come in contact with an object and cause some change in them".
Instrument:Measure	Describes an object or protocol that can measure another object(s).
Instrument:Observe	Describes an object which can be used to observe another object(s).
Instrument:Maintain	Describes an object or protocol which can be used to maintain the state of object(s).
Instrument:Catalyst	Describes an object that can be used as a catalytic "facilitator" for an experimental event to occur.
Instrument:Reference	Refers to a method or protocol being used.
Instrument:Mathematical	Describes a mathematical or computational instrument (e.g., simulation, algorithm, equation, and the use of software).

Table 2. Semantic Roles in the Annotation Scheme of our Experimental Event

includes: *Theme*, *Patient*, *Predicate*, *Agent*, *Location*, *Goal*, etc. The complete set of semantic roles and their definitions in our experimental event scheme is presented in Table 2.

We have extended the VerbNet definition of the semantic role *Instrument* from simply describing "an object or force that comes in contact with an object and causes some change in them" [22] to include a variety of subcategories that correspond to various types of biological and man-made instruments that are used in a biochemistry laboratory. Examples of these subcategories include:
1- Instruments used to *change* the state of an object. For example:

Example 2. "Beads with bound proteins were washed six times (for 10 min under rotation at 4 C) with **pulldown buffer** ..." [7].

In this example, the pulldown buffer was used to wash (change the state of) the Beads with bound proteins. In this instance, the phrase "pulldown buffer" should be labeled as **instrument (change)**.
2- Instruments used to *maintain* the state of an object. For example:

Example 3. "Once the samples were in EPR tubes, they were immediately frozen in liquid nitrogen, and stored in **liquid nitrogen** before using." [5].

In this example, the liquid nitrogen was used to store (maintain the condition of) the samples which were in the EPR tubes. In this case, the phrase "liquid nitrogen" should be labeled as **instrument (maintain)**.
3- Instruments used to *observe* an object. For example:

Example 4. The mitochondria was observed by **spinning disk confocal microscopy**.

The spinning disk confocal microscopy is used to observe the mitochondria. We should label the phrase "spinning disk confocal microscopy" as **instrument (observe)**.
4- Instruments used as a *catalyst* in experimental processes to occur. For example:

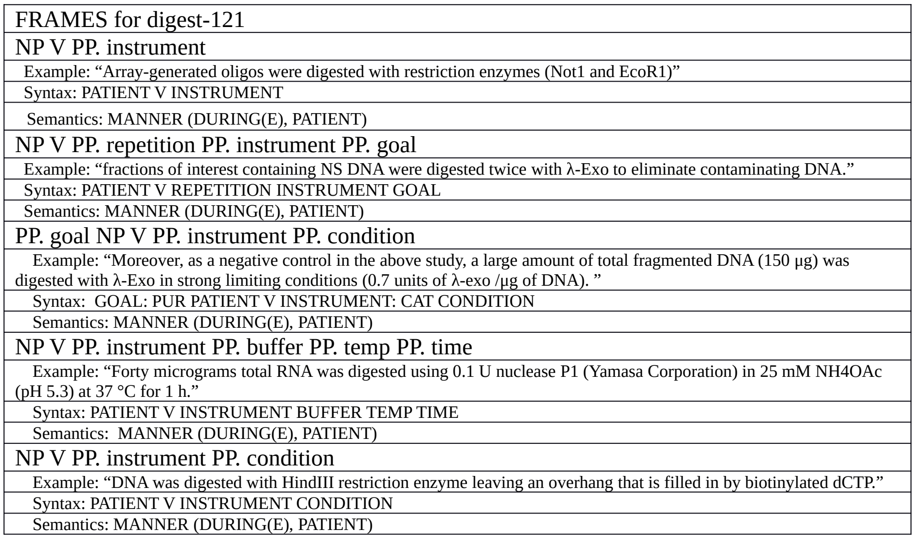
Example 5. "The ca. 900 bp PCR products were digested with **NdeI and HindIII** and ligated into pUC19." [4].

In this example, the NdeI and HindIII are enzymes used to facilitate the digestion (cutting) of the ca.(approximately) 900 bp PCR products. In this instance, the phrase "NdeI and HindIII" should be labeled as **instrument (catalyst)**.

We have also proposed a new semantic role *protocol detail* that identifies certain types of information about experimental processes. Examples of these subcategories include: 1- Time or the duration of a process [22]. For example:

Example 6. "Beads with bound proteins were washed six times (**for 10 min** under rotation at 4 C) with pulldown buffer ..." [7].

[9] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html
[9] http://www.glossary.sil.org/term/factitive-semantic-role



FRAMES for digest-121
NP V PP. instrument
Example: “Array-generated oligos were digested with restriction enzymes (Not1 and EcoR1)”
Syntax: PATIENT V INSTRUMENT
Semantics: MANNER (DURING(E), PATIENT)
NP V PP. repetition PP. instrument PP. goal
Example: “fractions of interest containing NS DNA were digested twice with λ-Exo to eliminate contaminating DNA.”
Syntax: PATIENT V REPETITION INSTRUMENT GOAL
Semantics: MANNER (DURING(E), PATIENT)
PP. goal NP V PP. instrument PP. condition
Example: “Moreover, as a negative control in the above study, a large amount of total fragmented DNA (150 μg) was digested with λ-Exo in strong limiting conditions (0.7 units of λ-exo /μg of DNA). ”
Syntax: GOAL: PUR PATIENT V INSTRUMENT: CAT CONDITION
Semantics: MANNER (DURING(E), PATIENT)
NP V PP. instrument PP. buffer PP. temp PP. time
Example: “Forty micrograms total RNA was digested using 0.1 U nuclease P1 (Yamasa Corporation) in 25 mM NH4OAc (pH 5.3) at 37 °C for 1 h.”
Syntax: PATIENT V INSTRUMENT BUFFER TEMP TIME
Semantics: MANNER (DURING(E), PATIENT)
NP V PP. instrument PP. condition
Example: “DNA was digested with HindIII restriction enzyme leaving an overhang that is filled in by biotinylated dCTP.”
Syntax: PATIENT V INSTRUMENT CONDITION
Semantics: MANNER (DURING(E), PATIENT)

Fig. 1. The verb frame for the verb *digest*

2- Temperature of an experimental process. For example:

Example 7. “Beads with bound proteins were washed six times (for 10 min under rotation **at 4 C**) with pulldown buffer ...” [7].

With these semantic roles we are able to provide the frames for procedural verbs. To illustrate, Fig. 1 contains the frame for the verb *digest*.

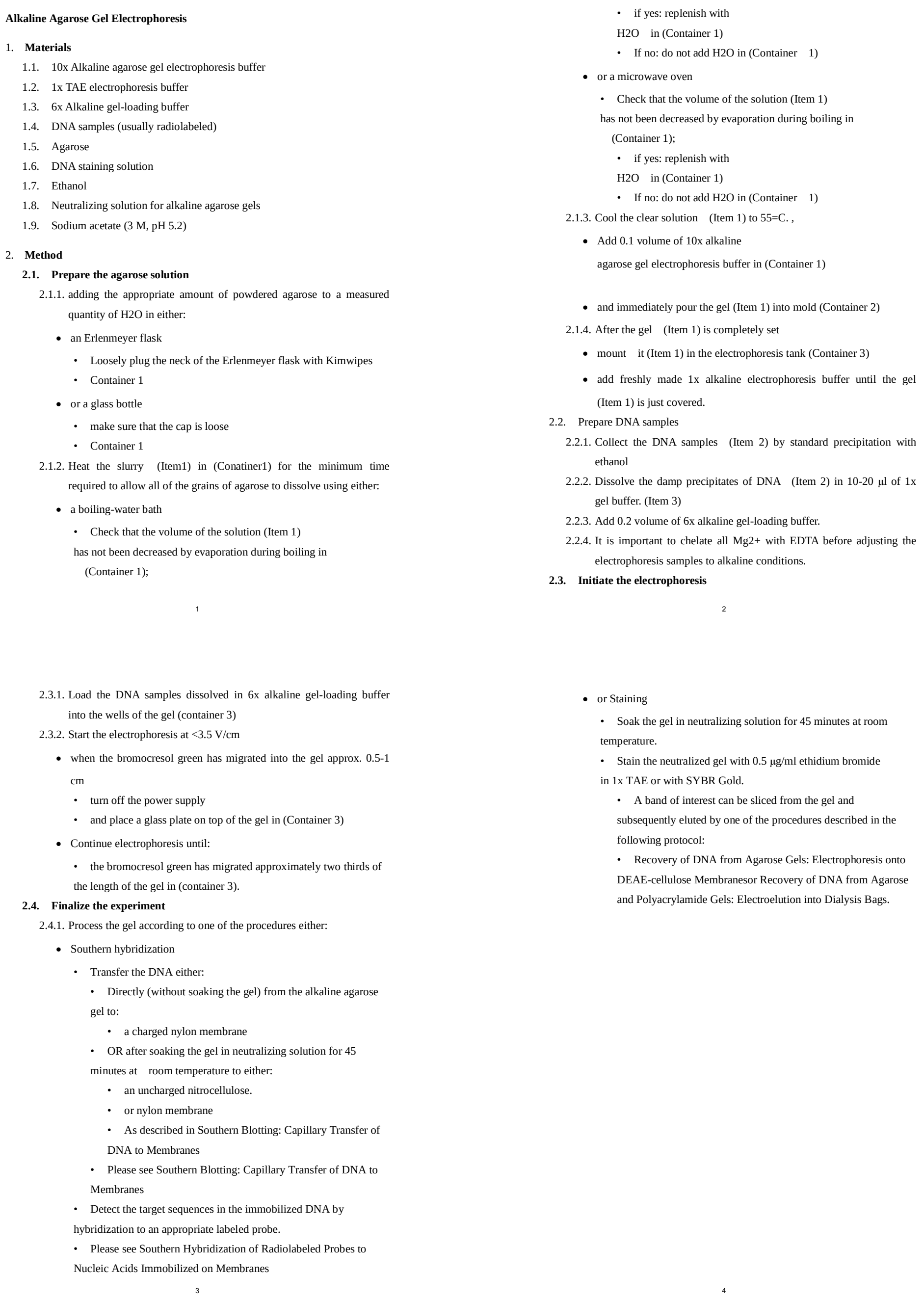
4 An Ontology of Biochemical Techniques and Laboratory Practices

Knowledge about how experiments are carried out in a biochemistry laboratory is absolutely essential to the understanding of much of the text found in biochemistry articles. We needed assistance from a biochemist to understand many of the sentences that are present in our corpus. With this in mind we have developed an ontology prototype to assist with a computational approach to analyzing the sentences found in the Methods section of a biochemistry article.

The example of a procedure called Alkaline Agarose Gel Electrophoresis is given in text format in Fig. 2. This is a common procedure used to isolate the biological substance that is used in future procedures from the other substances found in the solution that results from the previous procedures.

5 A Manual Annotation of a Portion of a Method Section

We have selected three articles from our corpus randomly to manually analyze and extract steps in experimental procedures (processes) from the method sec-



Alkaline Agarose Gel Electrophoresis

1. **Materials**
   - 1.1. 10x Alkaline agarose gel electrophoresis buffer
   - 1.2. 1x TAE electrophoresis buffer
   - 1.3. 6x Alkaline gel-loading buffer
   - 1.4. DNA samples (usually radiolabeled)
   - 1.5. Agarose
   - 1.6. DNA staining solution
   - 1.7. Ethanol
   - 1.8. Neutralizing solution for alkaline agarose gels
   - 1.9. Sodium acetate (3 M, pH 5.2)

2. **Method**
   - **2.1. Prepare the agarose solution**
     - 2.1.1. adding the appropriate amount of powdered agarose to a measured quantity of H2O in either:
       - an Erlenmeyer flask
         - Loosely plug the neck of the Erlenmeyer flask with Kimwipes
         - Container 1
       - or a glass bottle
         - make sure that the cap is loose
         - Container 1
     - 2.1.2. Heat the slurry (Item1) in (Conatiner1) for the minimum time required to allow all of the grains of agarose to dissolve using either:
       - a boiling-water bath
         - Check that the volume of the solution (Item 1) has not been decreased by evaporation during boiling in (Container 1);

   - if yes: replenish with H2O in (Container 1)
   - If no: do not add H2O in (Container 1)
 - or a microwave oven
   - Check that the volume of the solution (Item 1) has not been decreased by evaporation during boiling in (Container 1);
     - if yes: replenish with H2O in (Container 1)
     - If no: do not add H2O in (Container 1)
- 2.1.3. Cool the clear solution (Item 1) to 55=C. ,
  - Add 0.1 volume of 10x alkaline agarose gel electrophoresis buffer in (Container 1)
  - and immediately pour the gel (Item 1) into mold (Container 2)
- 2.1.4. After the gel (Item 1) is completely set
  - mount it (Item 1) in the electrophoresis tank (Container 3)
  - add freshly made 1x alkaline electrophoresis buffer until the gel (Item 1) is just covered.
- 2.2. Prepare DNA samples
  - 2.2.1. Collect the DNA samples (Item 2) by standard precipitation with ethanol
  - 2.2.2. Dissolve the damp precipitates of DNA (Item 2) in 10-20 µl of 1x gel buffer. (Item 3)
  - 2.2.3. Add 0.2 volume of 6x alkaline gel-loading buffer.
  - 2.2.4. It is important to chelate all Mg2+ with EDTA before adjusting the electrophoresis samples to alkaline conditions.
- **2.3. Initiate the electrophoresis**

- 2.3.1. Load the DNA samples dissolved in 6x alkaline gel-loading buffer into the wells of the gel (container 3)
- 2.3.2. Start the electrophoresis at <3.5 V/cm
  - when the bromocresol green has migrated into the gel approx. 0.5-1 cm
    - turn off the power supply
    - and place a glass plate on top of the gel in (Container 3)
  - Continue electrophoresis until:
    - the bromocresol green has migrated approximately two thirds of the length of the gel in (container 3).
- **2.4. Finalize the experiment**
  - 2.4.1. Process the gel according to one of the procedures either:
    - Southern hybridization
      - Transfer the DNA either:
        - Directly (without soaking the gel) from the alkaline agarose gel to:
          - a charged nylon membrane
        - OR after soaking the gel in neutralizing solution for 45 minutes at room temperature to either:
          - an uncharged nitrocellulose.
          - or nylon membrane
          - As described in Southern Blotting: Capillary Transfer of DNA to Membranes
        - Please see Southern Blotting: Capillary Transfer of DNA to Membranes
      - Detect the target sequences in the immobilized DNA by hybridization to an appropriate labeled probe.
      - Please see Southern Hybridization of Radiolabeled Probes to Nucleic Acids Immobilized on Membranes

- or Staining
  - Soak the gel in neutralizing solution for 45 minutes at room temperature.
  - Stain the neutralized gel with 0.5 µg/ml ethidium bromide in 1x TAE or with SYBR Gold.
    - A band of interest can be sliced from the gel and subsequently eluted by one of the procedures described in the following protocol:
    - Recovery of DNA from Agarose Gels: Electrophoresis onto DEAE-cellulose Membranesor Recovery of DNA from Agarose and Polyacrylamide Gels: Electroelution into Dialysis Bags.

Fig. 2. Alkaline Agarose Gel Electrophoresis Ontology

No.	Sentence
1	The over-expression plasmid for L1, pUB5832, was digested with *Nde*I and *Hind*III, and the resulting ca. 900 bp piece was gel purified and ligated using T4 ligase into pUC19, which was also digested with *Nde*I and *Hind*III, to yield the cloning plasmid pL1PUC19.
2	Mutations were introduced into the L1 gene by using the overlap extension method of Ho et al. [60], as described previously [68].
3	The oligonucleotides used for the preparation of the mutants are shown in Table 1.1.

Table 3. Some sentences from the article **Biochem-3-_-77373** [4]

tion. Table 3 shows some sentences from one of these articles [4]. The purpose of this analysis is to identify the semantic roles of experimental processes and the semantic frames of procedural verbs that occurred in these processes. Also, we want to demonstrate the usefulness of our approach by mapping the knowledge of frame semantics and the ontological knowledge to rhetorical moves.

The sentences in Table 3 are three contiguous sentences in a biochemistry article. They discuss the idea of cutting a DNA piece of a plasmid, which is "a small circular and double-stranded DNA molecule that is distinct from a cell's chromosomal DNA"[10], and ligate (attach) that piece to another plasmid to produce the desired protein. Table 4 shows five events from the sentences in Table 3. The events 1, 2, 3, and 4 are extracted from Sentence No. 1 and Sentence No. 2 has only Event 5, while there is no actual experimental event in Sentence No. 3. It rather simply refers to a table in the article's prior text. Each event in Table 4 represents one complete experimental procedure. Also the actual sequence of experimental events in the lab don't necessarily follow the sequence that these events appear in the text. Another important aspect to note is that not all the essential information about experimental processes is found in the text, some information can be implied. However, these implied pieces of information can be inferred from an ontology of standard biochemistry procedures, some of which we have developed. Taking a look at Events 1-4 in Table 4:
1- Digestion of pUB5832: a 900 bp piece was cut out using two restriction enzymes (*Nde*I and *Hind*III)
2- Then, the gel purification of the 900 bp piece: gel electrophoresis was used in this purification step. This is implied information derived from the ontology.
3- At any time before Event 4, the digestion of pUC19 happens, This could happen before, after, between, or during Events 1 and 2.
4- After Events 1, 2, and 3, ligation of the 900 bp into pUC19 occurs.

A lot of information can be derived from the text using knowledge about the verbs. This has been described earlier: the semantic roles of each verb together with syntactic information allows this information to be extracted from

[10] plasmid / plasmids — Learn Science at Scitable. (n.d.). Retrieved December 22, 2017, from https://www.nature.com/scitable/definition/plasmid-plasmids-28

Event 1	Event 2	Event 3
Sentence No. 1 – Patient: The over-expression plasmid for L1, pUB5832 – Predicate: digested – Instrument (catalyst): *Nde*I and *Hind*III	Sentence No. 1 – Patient: the resulting ca. 900 bp piece – Predicate: gel purified – Instrument (catalyst): Gel electrophoresis	Sentence No. 1 – Patient: pUC19 – Predicate: digested – Instrument (catalyst): NdeI and HindIII
Event 4	Event 5	
Sentence No. 1 – Patient: the resulting ca. 900 bp piece – Predicate: ligated – Instrument (catalyst): using T4 ligase – goal: into pUC19	Sentence No. 2 – Patient: the L1 gene – Predicate: introduced (mutated) – Instrument (reference type): using the overlap extension method of Ho et al.	Sentence No. 3 does not contain experimental events.

Table 4. Extracted events from two sentences in the article **Biochem-3-_-77373** [4]

the text. Table 4 shows this extracted information. However, this is not enough to understand the information provided in the text.

A proper interpretation of the description of events in Sentence No. 1 cannot be completely derived from the text alone. An understanding of laboratory practice together with knowledge of what is involved in performing plasmid digestion, purification, and ligation is required. Some of the event sequencing can be derived from the text, for instance, the pragmatics of the conjunction "and" usually indicates that the second conjunct follows temporally after the first conjunct has completed. The phrase "the resulting" is also a key linguistic clue to determine this sequence. But, when the third event happens requires knowledge of biochemistry and laboratory practice as well as knowledge of the complete method. The linguistic information provided by the use of a relative clause does not enable a complete understanding of this event, so the ontology is required for the information required to do a proper interpretation. Another important aspect of the text is that all of the referents are described by a singular nouns. However, knowing the biological processes that are carried out in the laboratory is important: solutions containing large numbers of the biological elements are used. Hence, one is not dealing with a single plasmid or a single piece of the plasmid, and when the digestion occurs, all of the pieces of the plasmids are in the solution including ones that didn't get digested, thus the need for the gel purification step which separates the various biological elements.

6 Conclusions and Future Work

In this research we have provided prototypes for two ontologies of the biochemistry domain. The first ontology, *procedurally rhetorical frame semantics*, provides semantic roles for procedural verbs. The second ontology provides information about biochemical techniques. This ontology can be used to give information that does not appear in the scientific article text. To the best of our knowledge, no research has proposed or incorporated the idea of a semantic frame based on verb analysis to assist in the analysis of argumentation in biochemistry articles. Nor has any attempt been made to build an ontology of biochemical techniques and laboratory practices.

Our future goal is an in-depth argumentation analysis of biochemistry articles. Having access to the rhetorical moves that have been extracted using the two ontologies will enable a computationally feasible technique that will enable argumentation mining of more-detailed scientific knowledge than is currently available. This will be an important step towards providing researchers in Computational Argumentation working in domains with similar discourse structure with a means of using and evaluating the metrics we will develop.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. pp. 86–90. Association for Computational Linguistics (1998)
2. Biber, D.: Variation across speech and writing. Cambridge University Press (1991)
3. Boyer, R.F.: Biochemistry Laboratory: Modern Theory and Techniques. Prentice Hall (2012)
4. Carenbauer, A.L., Garrity, J.D., Periyannan, G., Yates, R.B., Crowder, M.W.: Probing substrate binding to Metallo-β-Lactamase L1 from Stenotrophomonas maltophilia by using site-directed mutagenesis. BMC Biochemistry 3(1), 4 (Feb 2002), https://doi.org/10.1186/1471-2091-3-4
5. Chen, W., Guidotti, G.: The metal coordination of scd39 during atp hydrolysis. BMC biochemistry 2(1), 9 (2001)
6. Cohen, K.B., Demner-Fushman, D.: Biomedical natural language processing, vol. 11. John Benjamins Publishing Company (2014)
7. Ester, C., Uetz, P.: The ff domains of yeast u1 snrnp protein prp40 mediate interactions with luc7 and snu71. BMC biochemistry 9(1), 29 (2008)
8. Fillmore, C.J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences 280(1), 20–32 (1976)
9. Fillmore, C.J.: Topics in lexical semantics. Current issues in linguistic theory 76, 138 (1977)
10. Green, N.: Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In: Proceedings of the first workshop on argumentation mining. pp. 11–18 (2014)
11. Green, N.: Identifying argumentation schemes in genetics research articles. In: Proceedings of the 2nd Workshop on Argumentation Mining. pp. 12–21 (2015)

12. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics 17(1), 132–144 (2016)
13. Hunter, L., Cohen, K.B.: Biomedical language processing: What's beyond pubmed? Molecular Cell 21(5), 589–594 (2006)
14. Hur, J., Özgür, A., Xiang, Z., He, Y.: Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. Journal of Biomedical Semantics 3(1), 18 (2012)
15. Kanoksilapatham, B.: A corpus-based investigation of scientific research articles: Linking move analysis with multidimensional analysis. Ph.D. thesis, Georgetown University (2005)
16. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of biocreative ii. Genome biology 9(2), S4 (2008)
17. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28(7), 991–1000 (2012)
18. Lu, Z.: Pubmed and beyond: a survey of web tools for searching biomedical literature. Database 2011 (2011)
19. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles. Computational linguistics 31(1), 71–106 (2005)
20. Ravikumar, K., Liu, H., Cohn, J.D., Wall, M.E., Verspoor, K.: Literature mining of protein-residue associations with graph rules learned through distant supervision. Journal of biomedical semantics 3(3), S2 (2012)
21. Sambrook, J., Russell, D.W.: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press (2001)
22. Schuler, K.K.: Verbnet: a broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania (2005)
23. Segura-Bedmar, I., Martinez, P., de Pablo-Sánchez, C.: Extracting drug-drug interactions from biomedical texts. BMC bioinformatics 11(5), P9 (2010)
24. Swales, J.: Genre analysis: English in academic and research settings. Cambridge University Press (1990)
25. Thompson, P., Nawaz, R., McNaught, J., Ananiadou, S.: Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics 12(1), 393 (2011)