Complex Event Extraction using DRUM

James Allen[1,2] jallen@ihmc.us
Will de Beaumont[1] wbeaumont@ihmc.us
Lucian Galescu[1] lgalescu@ihmc.us
Choh Man Teng[1] cmteng@ihmc.us

[1]Institute for Human and Machine Cognition, 40 S. Alcaniz Street, Pensacola FL 32502, USA
[2]Department of Computer Science, University of Rochester, Rochester NY 14627, USA

Abstract

Complex mechanisms, such as cell-signaling pathways, consist of many highly interconnected components, yet they are often described in disconnected fragmentary ways. The goal of DRUM (Deep Reader for Understanding Mechanisms) is to develop a system that can read papers and combine results of individual studies into a comprehensive explanatory model. A first step is to automatically extract relevant events and event relationships from the literature. This paper describes initial steps in extending an existing general deep language understanding system, TRIPS, to read biomedical papers. In a preliminary evaluation, our system was the best performing system among the participants, achieving results close to human expert performance. These results suggested that our system is viable for complex event extraction and, ultimately, understanding complex systems and mechanisms.

1. Introduction

Complex mechanisms consist of many highly interconnected components, yet they are often described in disconnected fragmentary ways. Examples include ecosystems, social dynamics and signaling networks in biology. The study of these complex systems is often focused on a small portion of a mechanism at a time. In addition, the huge volume of scientific literature makes it difficult to track the fast developments in the field to achieve a comprehensive understanding of the often distant and convoluted interactions in the system.

The goal of the DRUM (Deep Reader for Understanding Mechanisms) project is to develop a system that can read papers and combine research results of individual studies into a comprehensive explanatory model of a complex mechanism. The system will automatically read scientific papers, extract relevant new model fragments, and compose them into larger models that will expose the interactions and relationships between disparate elements in the mechanism.

A first step towards this goal is to automatically extract relevant events and event relationships from the literature. In this paper we will describe initial steps in extending an existing general deep language understanding system, TRIPS (Allen et al, 2008), to the genre of scientific writing, in particular in the biomedical domain. Events in biomedical research papers are described in a highly specialized and technical language, with complex formulations and nested constructions. We will discuss adaptations made and how the design principles of TRIPS facilitate such adaptations.

We will report on an experimental evaluation of this extended system on extracting events and relationships centered on the Ras signaling pathways from a number of text passages in scientific papers. Our system was the best performing system among those evaluated, achieving results close to human expert performance.

Admittedly this was a small and preliminary evaluation. However, the results suggested our system is viable for complex event extraction. Of note, unlike typical statistical approaches, we did not train on text describing the Ras signaling pathways (or on any other text for that matter). Our results were achieved using a general deep language understanding system, with little domain-specific customization beyond the recognition of named entities and some specialized vocabularies. Most important, our goal does not stop at the surface extraction of events, as is the case for many existing bio-event extraction tasks. With a general deep language understanding system, we are in a good position to develop an *understanding* of the underlying connections in complex models, and the methods developed to achieve that understanding will be readily transferrable to domains other than biology.

2. The TRIPS Architecture

Much recent text processing work has focussed on developing “shallow”, statistically driven techniques. TRIPS takes a different approach, using statistical methods as a preprocessing step to provide guidance to a deep parsing system that uses a detailed, hand-built, grammar of English with a rich set of semantic restrictions. Figure 1 shows an overview of the system architecture. In

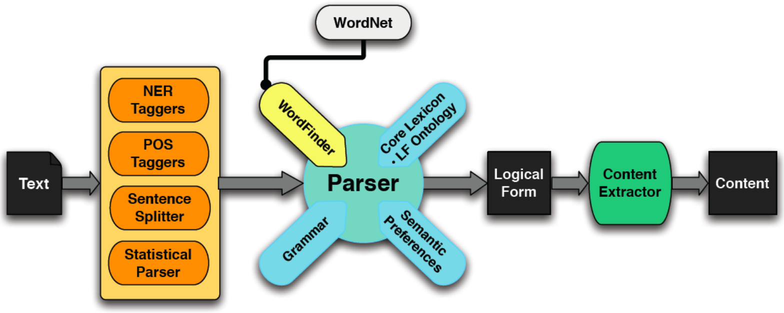


Figure 1. System Architecture.

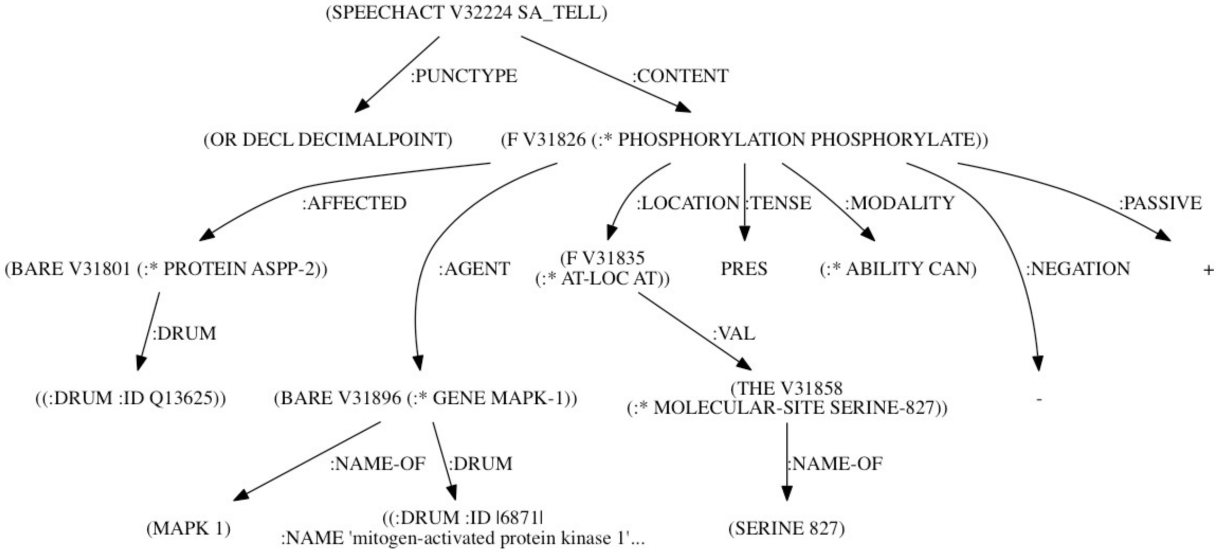

(SPEECHACT V32224 SA_TELL)
:PUNCTYPE
:CONTENT
(OR DECL DECIMALPOINT)
(F V31826 (:* PHOSPHORYLATION PHOSPHORYLATE))
:AFFECTED
:AGENT
:LOCATION
:TENSE
:MODALITY
:NEGATION
:PASSIVE
(BARE V31801 (:* PROTEIN ASPP-2))
(F V31835 (:* AT-LOC AT))
PRES
(:* ABILITY CAN)
+
:DRUM
:VAL
((:DRUM :ID Q13625))
(BARE V31896 (:* GENE MAPK-1))
(THE V31858 (:* MOLECULAR-SITE SERINE-827))
-
:NAME-OF
:DRUM
:NAME-OF
(MAPK 1)
((:DRUM :ID |6871|
:NAME 'mitogen-activated protein kinase 1'...
(SERINE 827)

Figure 2. The logical form produced by DRUM of the sentence "*ASPP2 can be phosphorylated at serine 827 by MAPK1.*"

the rest of this section we will describe briefly the main components of the system. The content extractor, customized for biomedical text, will be discussed in more detail in Section 4.

2.1. Parser

The TRIPS grammar is a lexicalized context-free grammar, augmented with feature structures and feature unification. The grammar is motivated by X-bar theory (Jackendoff, 1977), and draws on principles from GPSG (Gazdar et al., 1985), for example head and foot features, and HPSG (Pollard and Sag, 1987, 1994). The search in the parser is controlled by a set of hand-built preferences encoded as weights on the rules, together with domain-general selectional restrictions (encoded in the lexicon and ontology) to eliminate semantically anomalous sense combinations.

The TRIPS parser uses a packed-forest chart representation and builds constituents bottom-up using a best-first search strategy similar to A*, based on rule and lexical weights and the influences of the front end components (described below).

The parser constructs from the input a logical form, which is a semantic representation that captures an unscoped modal logic (Allen, 1995; Manshadi et al., 2008). The logical form includes the surface speech act, semantic types, semantic roles for predicate arguments, and dependency relations. Consider the sentence:

> *ASPP2 can be phosphorylated at serine 827 by MAPK1.*

Figure 2 is a graphical depiction of the logical form of this sentence produced by DRUM. The nodes in the graph represent the word senses and ontology types, together with quantification information, and the edges represent semantic role relations. Of particular interest are two of the core semantic roles: AGENT (here, *MAPK1*), identifying objects that play a causal role in an event; and AFFECTED (here, *ASPP2*), identifying objects that are changed as part of an event. Other roles also provide key information that needs to be extracted. For instance, LOCATION identifies the molecular site (here, *serine 827*) or cellular location (e.g., *cytoplasm*) associated with an event of interest. The logical form also captures tense, modality and aspect information, which is crucial for determining, for example, whether a statement is a stated fact, a conjecture or a possibility (as indicated by the modality).

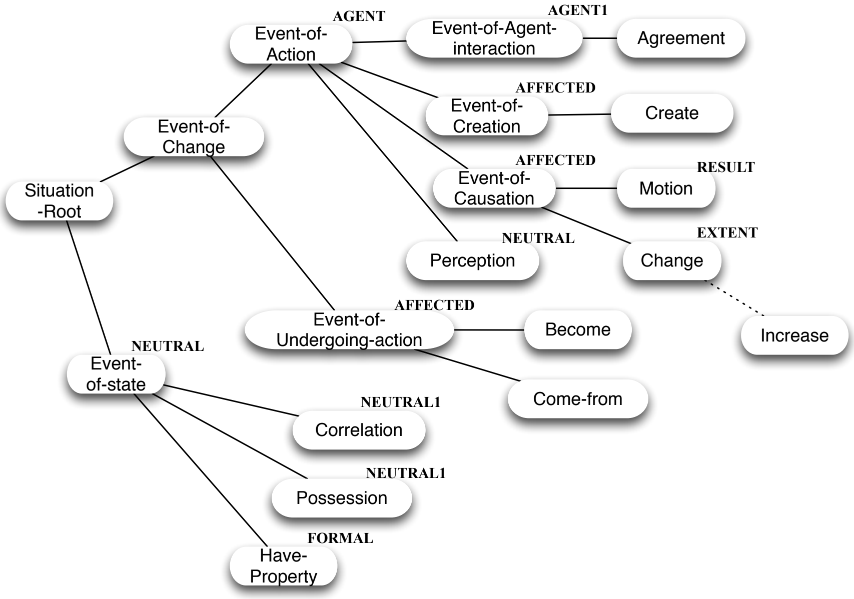


Figure 3. A subset of the TRIPS upper event ontology, showing core roles

2.2. Ontology and Lexicon

The parser draws on a general purpose semantic lexicon and ontology which define a range of word senses and lexical semantic relations. The core semantic lexicon was constructed by hand and contains approximately 7,500 lemmas (generating approximately three times that many words) and 2,000 concepts in the ontology.

The ontology is organized hierarchically and each ontology concept has associated with it possible semantic roles and selectional preferences that further refine the concept. For instance, it can be specified that the AFFECTED role of *phosphorylate* may only take a physical object that is part of a molecule (e.g., a protein or a molecular site). Figure 3 shows a portion of the TRIPS upper ontology for events and their associated core semantic roles.

A TRIPS lexicon entry is composed of two key parts. The first is the ontology type of the word sense, and it receives the roles and restrictions specific to its ontology type together with those inherited from its ancestors in the ontology hierarchy. The second is the grammatical constructions that the word can participate in, in the form of rules that map syntactic patterns to instantiations of objects from the ontology.

2.3. Extending the Lexicon

To attain broad lexical coverage beyond its hand-defined core lexicon, TRIPS uses input from a variety of external resources, some of which will be described in the next sections. Using the built-in subsystem *WordFinder,* TRIPS can augment its lexicon by dynamically building lexical entries with plausible semantic and syntactic structures for virtually any word in WordNet (Fellbaum, 1998), thus extending its coverage to over 100,000 words.

For words not in the core lexicon, WordFinder uses a hand-built mapping between the hypernym information in WordNet (for all the WordNet senses) and the TRIPS ontology. For each identified TRIPS class it gathers all the possible constructions that words of this class in the TRIPS lexicon participate in. It then generates a set of lexical entries for the unknown word by combining each possible ontological class with each possible construction for that class. While this procedure may over-generate, the key is to include the correct constructions among the generated possibilities, since these correct constructions will be the ones realized in parsing sentences (for more information see Allen, 2014).

2.4. Front End Components

To support more robust processing and domain configurability, the core system has the capability to incorporate a variety of statistical and symbolic natural language processing components in the front end, as well as domain-specific components such as specialized named entity recognizers. These include several off-the-shelf natural language tools such as the Shlomo Yona sentencizer[1], the Stanford part-of-speech tagger (Toutanova and Manning, 2000), the Stanford named-entity recognizer (NER) (Finkel et al., 2005) and the Stanford Parser (Klein and Manning, 2003). The output of these and other spe-

[1] http://search.cpan.org/~kimryan/Lingua-EN-Sentence-0.29/

Resource	Entities	References
BRENDA Tissue Ontology	tissues, cell types, cell lines	Gremse et al., 2011
Cell Ontology (CL)	cell types	Diehl et al., 2011
Chemical Entities of Biological Interest (ChEBI)	chemicals, molecule types, cell components	Degtyarenko et al., 2008
Gene Ontology (GO)	molecular functions, biological processes, pathways, cell components, macromolecular complexes	Ashburner et al., 2000
HUGO Gene Nomenclature (HGNC)	genes	Gray et al., 2015
Medical Subject Headings (MeSH®), Supplementary Concept Records (SCR)	drugs and chemicals	Lipscomb, 2000
neXtProt	cell lines, protein families	Gaudet et al., 2015
Pfam	protein families	Finn et al., 2014
Proteomics Standards Initiative for Molecular Interaction (PSI-MI)	molecular interactions, molecule type, macromolecular complexes, genes and proteins, biological roles, units of measurement	Hermjakob et al., 2004
UniProtKB (Swiss-Prot)	proteins	Uniprot Consortium, 2014

Table 4. Sources of domain-specific terminology/concepts and the types of entities incorporated into the TRIPS ontology

cialized preprocessors (e.g., a street address recognizer) is sent to the parser as advice. The parser then decides whether to follow these pieces of advice as it searches for the optimal parse of the sentence.

3. Extensions and Customization for the Biomedical Genre

We describe below several extensions to the general TRIPS system to better handle the text characteristics of the biomedical literature.

3.1. Genre Specialization

The chart produced by the parser is searched using a dynamic programming algorithm to find the least cost sequence of constituents according to a cost table that can be varied by genre. For instance, in dialogue systems speech acts such as CONFIRM (e.g., ok) or GREET (e.g., hello) are expected. For papers in the biomedical domain, such speech acts almost never occur and thus are discounted in favor of TELL statements. Similarly, in dialogue systems utterances are expected to be fairly short and colloquial, whereas in scientific text the sentence structures are expected to be much more formal and involved. The parameters for parsing and the cost table are set accordingly.

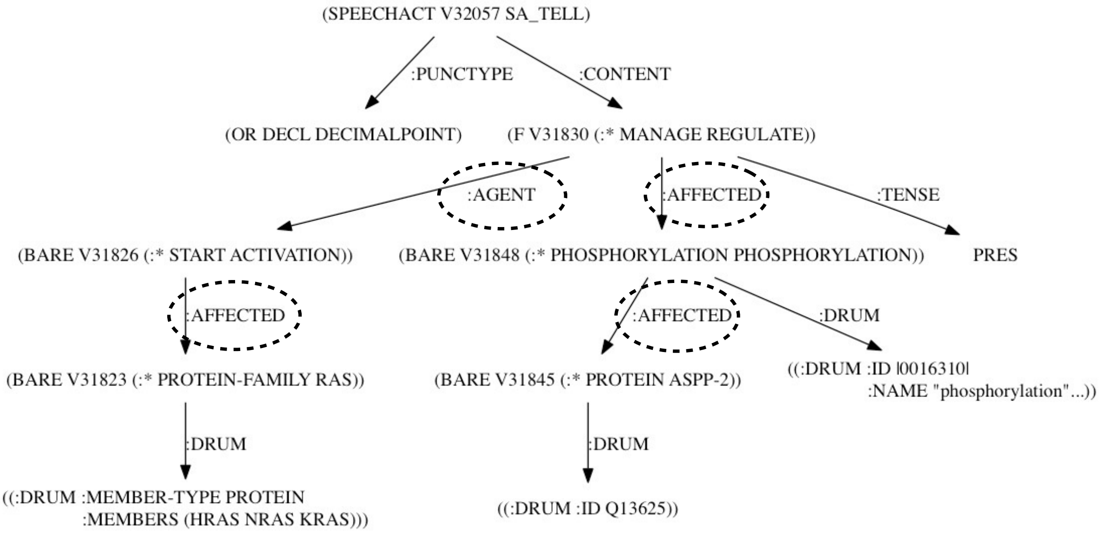
In addition, the system can choose to incorporate different front end components. For instance, for the biomedical literature a street address recognizer would not be very useful, but a named entity recognizer for protein names would be most crucial.

Such customizations not only optimize the parser efficiency, but also reduce the potential ambiguities the parser has to deal with, since each additional component offers additional, potentially conflicting, advice the parser has to take into account.

3.2. Lexicon and Ontology Enhancements

The biomedical domain uses specific terminology that is outside the core TRIPS lexicon and ontology. We extended the system's coverage by incorporating domain-specific terminology, with mappings to TRIPS ontology classes. In some cases we introduced new ontology categories to accommodate domain-specific concepts. Table 4 lists the resources used, as well as the types of entities mapped to the TRIPS ontology. Some of these resources organize concepts in ontologies (e.g., using the OBO format (Smith et al., 2007)); for these, we grafted the relevant nodes onto the TRIPS ontology (see Blaylock et al., 2011). For example, most GO biological processes are mapped to the existing TRIPS ontology category ONT::event-of-change; however, children of GO:0007165 (signal transduction) are names/types of signaling pathways, and they are mapped to ONT::signaling-pathway—a domain-specific category newly added to the TRIPS ontology. Controlled vocabularies for single entity types (e.g., neXtProt's Cellosaurus) were mapped to single TRIPS ontology types (e.g., ONT::cell-line).

In addition, we used the SPECIALIST lexicon (McCray et al., 1994) for obtaining syntactic category information about domain-specific lexical items, which is helpful during parsing; however, since SPECIALIST does not include semantic information, the lexical entries are not mapped into the TRIPS ontology.


(SPEECHACT V32057 SA_TELL)
:PUNCTYPE
:CONTENT
(OR DECL DECIMALPOINT)
(F V31830 (:* MANAGE REGULATE))
:AGENT
:AFFECTED
:TENSE
(BARE V31826 (:* START ACTIVATION))
(BARE V31848 (:* PHOSPHORYLATION PHOSPHORYLATION))
PRES
:AFFECTED
:AFFECTED
:DRUM
(BARE V31823 (:* PROTEIN-FAMILY RAS))
(BARE V31845 (:* PROTEIN ASPP-2))
((:DRUM :ID |0016310| :NAME "phosphorylation"...))
:DRUM
:DRUM
((:DRUM :MEMBER-TYPE PROTEIN :MEMBERS (HRAS NRAS KRAS)))
((:DRUM :ID Q13625))

```
(EVENT ONT::V31830 ONT::REGULATE :AGENT ONT::V31826 :AFFECTED ONT::V31848 :TENSE W::PRES
   :START 0 :END 46)

(EVENT ONT::V31826 ONT::ACTIVATE :AFFECTED ONT::V31823 :START 0 :END 15)

(EVENT ONT::V31848 ONT::PHOSPHORYLATION :AFFECTED ONT::V31845 :DRUM ((:DRUM :ID GO::|0016310|
   :NAME "phosphorylation"...)) :START 25 :END 46)

 (TERM ONT::V31823 ONT::PROTEIN-FAMILY :NAME W::RAS :DRUM ((:DRUM :MEMBER-TYPE ONT::PROTEIN
   :MEMBERS (HRAS NRAS KRAS))) :START 0 :END 4)

(TERM ONT::V31845 ONT::PROTEIN :NAME W::ASPP-2 :DRUM ((:DRUM :ID UP::Q13625)) :START 25 :END 31
```

Figure 5. The logical form of "*RAS activation regulates ASPP2 phosphorylation.*" and the events and terms extracted by DRUM.

3.3. Specialized Constructions

The TRIPS component *WordFinder* can construct lexical entries for words not explicitly found in the core lexicon, using a mapping between WordNet and the TRIPS ontology. This mechanism provides broad coverage of words in general use. However, certain "everyday" words have specialized usage in biology. For instance, "association" is not just a vague relationship but a specific kind of binding between molecules. Some other words are used in idiosyncratic constructions. For instance, "the protein localizes to the nucleus", which means the protein exists in the nucleus, required a novel syntactic template (and semantic characterization). These words pose particular difficulties for our system as our automatically derived general constructions would be inadequate. For such cases we often have to provide hand-tailored lexical entries with appropriate syntactic templates and semantic restrictions to distinguish the everyday and biological senses of the words.

3.4. Handling Nominalizations

Nominalization is prevalent in the biomedical genre (see for instance the example sentence in Figure 5). The TRIPS parser has a general mechanism for handling verb nominalizations. This is enabled by the fact that the ontological information is identical between the verbal and nominal forms of the same event (e.g., *dominate* and *dominance*). The only difference between verbal and nominal forms is the grammatical linking rules involved. For instance, for verbal forms the subject of a certain verb might map to the AGENT role, and the direct object to the AFFECTED role. In nominalizations, the possessive would map to the role identified with the subject of the verbal form, and the object of an *of* prepositional phrase would map to the role identified with the direct object of the verb. While there are a number of different constructions used with nominals, they appear to be generic across the entire set of nominalizations, and a set of a dozen or so generic rules is all that is needed. In addition, virtually all adjunct modifications (e.g., *for three hours*) apply equally well to both verbal and nominal forms using the same adverbial modification rules in the grammar.

4. Event Extraction

From the logical forms produced by the extended TRIPS parser we need to extract the events and event relationships of interest. Because much of the variation expected in sentence constructions is handled by the extended TRIPS system, we are

```
rule-activate (40): ACTIVATE(AGENT, AFFECTED) ← [ONT::start ONT::start-object] (AGENT, AFFECTED)
rule-decrease (20): DECREASE(AGENT, AFFECTED) ← [ONT::decrease] (AGENT, AFFECTED)
```

Figure 6. Specification of the extraction rules for two event types

able to use a relatively compact specification for defining the events and relationships of interest, while coping with fairly complex and nested formulations.

4.1. An Example

Consider the sentence:

> *RAS activation regulates ASPP2 phosphorylation.*

whose logical form is depicted in Figure 5. There are three events in this sentence: the central *regulation* event and two nested events, *activation* and *phosphorylation,* that serve as the arguments to the *regulation* event. The extractions of the three events are also shown in Figure 5, together with the two terms, *RAS* and *ASPP2*, involved in the events. Note that the word "activation" is mapped to the TRIPS ontology type ONT::start. It is this ontology type that triggers the extraction rule for an ACTIVATE event (see Figure 6).

In addition to the AGENT and AFFECTED roles, the :DRUM slot provides DRUM-specific grounding information about the events and entities, mostly derived from bio-resources (see Section 3.2). For example, UP::Q13625 is the UniProt identifier for the protein *ASPP2*.

4.2. Extraction Rule Specification

We capitalized on the TRIPS ontology and parser to develop a compact and easy-to-maintain specification of event extraction rules. Instead of having to write one rule to match each keyword/ phrase that could signify an event, many of these words/phrases have already been systematically mapped to a few types in the TRIPS ontology, using a combination of the TRIPS internal lexicon and the WordFinder component which allows us to attain the coverage of WordNet. For instance, *accumulate, gain, amplify, multiply, boost, double*, among others, all map to the TRIPS ontology type ONT::increase.

In addition, the parser handles various surface structures, and the logical form returned contains normalized semantic roles. For example,

RAS activates RAF
RAF is activated by RAS
The activation of RAF by RAS
Activated RAF
RAF activation

all are parsed into the same basic logical form with the semantic roles AFFECTED: *RAF* and, where applicable, AGENT: *RAS*. Thus, we needed very few (often only one) extraction rule specifications for each event type, covering a wide range of words and syntactic patterns.

Finally, since most events of interest are events of action, the usage patterns of these event words are often essentially identical, modulo the ontology types that signify the events and (less often) the semantic roles that correspond to the event arguments. We generated these rules using a module with standardized rule components, parameterized by only the event-specific ontology types and semantic role mappings. For example, *X activates / decreases / regulates / phosphorylates Y,* though denoting different events, all exhibit the same basic structure with the main semantic roles AGENT and AFFECTED. Complements denoting for example molecular sites and cellular locations for the most part retain the same structure across event types.

Figure 6 shows the stylized specification of two event types, ACTIVATE and DECREASE. The ACTIVATE line is read as follows:

name of rule: activate
priority of rule: 40
name of event to be extracted: ACTIVATE
semantic role 1: AGENT
semantic role 2: AFFECTED
ontology types: ONT::start; ONT::start-object

where the rule priority determines which rule is selected when multiple rules apply, and the ontology types are those in TRIPS that map to the target event type (here, ACTIVATE). The semantic roles may have further constraints on the types that can fill these roles. For instance, only molecular and cellular participants (e.g., proteins, chemicals, nucleus) are of interest in the context of biological events.

Note the similarity between the information for ACTIVATE and DECREASE. The only difference between the two lines is the ontology types that represent the respective event types (ONT::start, ONT::start-object for the former and ONT::decrease for the latter). This compact representation makes it easy to specify, maintain and update the extraction rules.

These rules were developed from general principles rather than based on specific training samples on the Ras signaling pathways. They were subsequently augmented as we learned more about specific biological usages. Although we do base our rules on the biological literature, we emphasize that neither the extraction rules described above nor any of the domain-specific

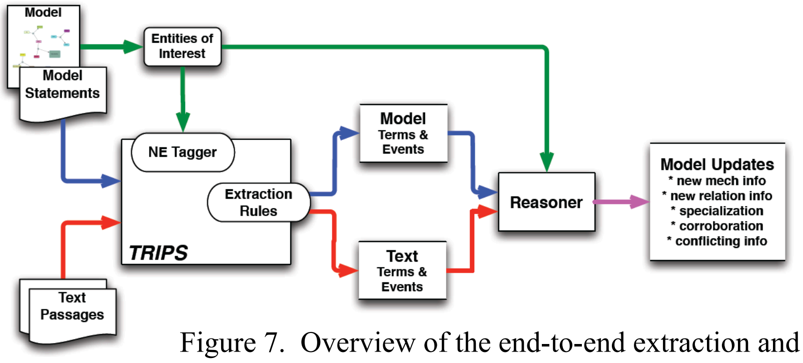


Figure 7. Overview of the end-to-end extraction and reasoning system.

> *We and others have recently shown that ASPP2 can potentiate RAS signaling by binding directly via the ASPP2 N-terminus [2,6]. Moreover, the RAS-ASPP interaction enhances the transcription function of p53 in cancer cells [2]. Until now, it has been unclear how RAS could affect ASPP2 to enhance p53 function. We show here that ASPP2 is phosphorylated by the RAS/Raf/MAPK pathway and that this phosphorylation leads to its increased translocation to the cytosol/nucleus and increased binding to p53, providing an explanation of how RAS can activate p53 pro-apoptotic functions (Figure 5). Additionally, RAS/Raf/MAPK pathway activation stabilizes ASPP2 protein, although the underlying mechanism remains to be investigated.*

Figure 8. Example text passage for evaluation.

enhancements to our system discussed in Section 3 are specific to the language or mechanisms describing the Ras signaling pathways. Thus, we expect our system to have comparable performance on any input describing bio-molecular mechanisms.

5. System Evaluation

We participated in a preliminary evaluation of event extraction, in the context of "reading with a model". A biological model was given in BioPAX (Demir et al., 2010), BEL (Selventa, 2011), and English. Given a set of text passages from scientific papers on the Ras signaling pathways, the goal was to extract from these passages events (and their arguments) that were relevant to the given model and make explicit the links between the extracted events and the model.

BioPAX and BEL do not have the linguistically motivated features and expressivity needed for our approach. To minimize hand coding and to create a uniform system, we created our initial model by reading and processing sentences simplified from the given English model sentences, using the same process as for reading and extracting information from the test passages. The model entities and events such processed were then compared to the entities and events extracted from the text passages. Figure 7 shows an overview of the automated end-to-end extraction and reasoning system.

Two types of events were distinguished here: mechanistic (e.g., X binds to Y) and regulatory/causal relationship (e.g., X increases Y). These were further classified with respect to the given model as: 1) new mechanism and 2) new relationship not in the model; 3) specialization and 4) corroboration of information in the model; and 5) conflict with the model. In addition, each result was to be accompanied by the supporting source text.

The reasoner aligned the extracted entities using their standardized identifiers (e.g., UniProt, HUGO, Gene Ontology). In addition, we derived the relationships between the model and text extractions based on the hierarchical organization of the event types. For instance, a regulation event subsumes a stimulation event, and thus "X regulates Y" corroborates "X stimulates Y" and the latter is a specialization of the former.

6. Results

Several passages, mainly from the results and discussion sections of two scientific papers, were selected as evaluation inputs. An example passage, from (Godin-Heymann et al., 2013), is shown in Figure 8.

The extractions and model comparisons were manually scored by a third party, based on the combined answers provided by two separate teams of biologists (30 events) and the addition of 5 events adopted from system submissions (see below). In "lenient" scoring for precision, incomplete results and results that were correct but irrelevant were excluded, whereas in "strict" scoring these results were counted as incorrect.

Eleven systems of varying degrees of automation participated in the evaluation. We have available only the lenient scores of other teams, as shown in Figure 9. For lenient scoring our system was the best performing system and our performance was close to human performance.

Note that although the human biologists had high precision, there was considerable non-overlap between the answers they provided. This accounted for the approximately 0.50 recall for either of the human teams, using the pooled answers of the two teams as the gold standard.

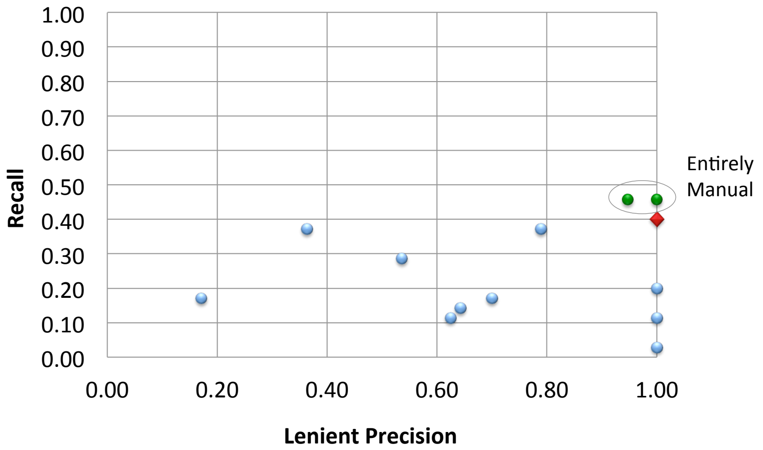


Figure 9. Evaluation results for eleven teams. The diamond ◆ represents the results of our system. The two topmost points are the manual scores of the two teams of human biologists.

Our precision, recall and F1 results for both the lenient and strict scorings are as follows:

	P	R	F1
lenient (strict)	1.00 (0.67)	0.40	0.57

7. Analysis

We believe precision is much more important than recall. A high precision system can generate valuable knowledge nuggets, even if it does not have high throughput, whereas output from a system with high recall but low precision cannot be trusted to be accurate. This is especially the case for such information-rich domains. Because of the huge volume of scientific literature, information is likely to be duplicated in multiple papers, and often also repeated in different forms in the same paper. Therefore, extracting (accurately) even a relatively small portion of the information in these papers could amount to a fair body of knowledge, even if we cannot extract everything from every sentence.

Our system showed promising performance on the evaluation data set. We achieved perfect precision, and recall close to the human experts. The modest recall even for the human experts indicated that this is a fairly difficult domain and there is not a clear-cut way to extract and encode the knowledge represented in these papers. In fact, after considering the submitted results, several additional events extracted by the systems but not by the human experts were incorporated into the gold standard.

We were able to extract some fairly complex, nested, events, similar to the one depicted in Figure 5. The ontology-based extraction and the lexical coverage extended by *WordFinder* allowed us to cope with a variety of expressions. For instance, from "... *ASPP2 can potentiate RAS signaling...*" we were able to map "potentiate" to an INCREASE event even though "potentiate" is not in the TRIPS core lexicon.

Another interesting example is "... *monoubiquitination abrogates GAP-mediated GTP hydrolysis*". This fairly complex sentence illustrates some of the strengths and weaknesses of our system. The system was able to extract two interleaving events:

ev1: REGULATE(AGENT: GAP; AFFECTED: ev2)
ev2: HYDROLYSIS(AFFECTED: GTP)

In the raw processing we also had the following:

ev3: INHIBIT(AGENT: MONOUBIQUITINATION; AFFECTED ev2)

but we failed to identify *what* was being monoubiquitinated and thus were not able to include this extraction in our results. The answer, that *Ras* was being monoubiquitinated, could only be identified with more sophisticated discourse processing.

We identified several main reasons for omissions in our extractions: 1) fragmented parses due to the long and complex sentence structures common in scientific publications; 2) insufficient domain-specific background knowledge, including language patterns specific to biology; 3) need for improved discourse processing and coreference resolution; and 4) lack of inference capabilities and persistent memory of inferences made.

The last point can be illustrated by the sentence "... *the RAS-ASPP interaction enhances the transcription function of p53...*" Here we need to be able to deduce that RAS-ASPP interaction produces a complex of the two, which then participates in further reactions.

As a final example, to be able to make sense of the seemingly simple sentence "*We obtained similar results using K-Ras...*" we need to address all of the above issues. Due to space limitation we will not discuss here the ongoing work towards tackling these challenges.

8. Related Work and Discussion

With the advent of relatively successful text mining strategies (named entity recognition, information extraction and retrieval) for the recognition and normalization of biologically relevant entities, automatic extraction of more complex, relational information from the biomedical literature has become a very active area of research. Shared Tasks (STs) such as the Protein-Protein Interaction (PPI) Task introduced at BioCreative II (Krallinger et al., 2008) and the BioNLP GENIA Event Extraction Task (Kim et al., 2009; Kim et al., 2011; Kim et al., 2013) have spurred a lot of activity in this area, although examples of earlier work certainly exist.

The goal in the PPI task is to extract binary protein-protein interaction pairs from full-text articles. More general biological events (e.g., regulatory events) beyond PPI involve much more varied relationships between entities and, indeed, between events themselves, leading to complex nested structures. The BioNLP STs have evolved to include more complex types of events and arguments. The GENIA ST (in particular 2013 which included coreference) and the Epigenetics and Post-translational Modifications task (EPI) introduced in 2011 (Ohta et al., 2011) are similar to our task. However, there are significant differences, too. We were not provided with gold annotations for entities; all relevant entities (including drugs, cell lines, cell components, sites) had to be extracted, and most of them had to be grounded in a reference database. Protein families were also important, as was the relation between families and the member proteins. Not only were coreferences supposed to be resolved, but, as indicated in Section 5, sometimes complex inferences were required to obtain a target event. In summary, our task was not designed to accommodate specific Information Extraction (IE) techniques; rather, in our evaluation the gold standard was human performance.

We would like to stress that our goal goes beyond IE. The need for deeper semantic approaches has been recognized before (see, e.g., Ananiadou et al., 2010). Still, the field is dominated by ML classifiers (for a list of the top-performing systems in the three BioNLP STs held so far, see Ananiadou et al., 2014). This sometimes results in seemingly paradoxical results, where systems can extract with relatively good performance phosphorylation events, but not ubiquitination events because the training data did not contain enough examples of the latter (Kim et al., 2011).

Indeed, ontological information is rarely used in current systems. GenIE (Cimiano et al., 2005) is an early example of an ambitious ontology-driven system that attempts to identify events based on constructing a full semantic representation of the text (using a semantic lexicon and semantic restrictions), as well as relations between events (using discourse information). The ontology they used, however, was a small, domain-specific one. To our knowledge the system has not been tested on any of the more recent event extraction tasks.

Although semantic (deep) parsing techniques have been rarely used for bio-event extraction, we note the PPI extraction study by Miyao et al. (2009), who found an HPSG-based parser to outperform (particularly in terms of precision) dependency and syntactic parsers, especially when trained on domain-specific corpora. However, they used the predicate-argument structures output by the parser as additional features for a statistical classifier.

In contrast, we do not depend on training with a domain specific corpus (although we have the capability to incorporate modules that do); rather, we extract events directly from the predicate-argument structures represented in the logical form, based on linguistic first principles that can be easily adapted to different domains. The advantage of this approach can be readily seen in this evaluation, in which, with a relatively short (but intensive) ramp up, we were able to outperform all other systems in the extraction of complex events and event relations. Of note, this was despite the fact that our system had lower named entity recognition scores than most others, particularly those with a history of participation in biomedical information extraction shared tasks.

The purpose of this evaluation was not a rigorous ranking of the different participating systems. Rather, we learned key areas we needed to improve. The results of this evaluation suggested that our system is viable for complex event extraction. This is however only the first step in understanding complex models and mechanisms. A general deep language understanding system that can be extended with domain-specific information will allow us to go beyond standard surface extraction tasks and develop the capabilities to truly *understand* big and complex mechanisms.

Acknowledgement

We thank the anonymous reviewers for comments that made this paper better. This work has been partially supported by the DARPA Big Mechanism program under ARO grant W911NF-14-1-0391.

References

Allen, J. F. (1995). Natural Language Understanding. Redwood City, CA, Benjamin Cummings.

Allen, J., M. Swift, et al. (2008). Deep Semantic Analysis of Text. Symposium on Semantics in Systems for Text Processing (STEP), Venice, Italy.

Allen, J. F. (2014). Learning a Lexicon for Broad-coverage Semantic Parsing. ACL Workshop on Semantic Parsing. Baltimore, MD.

Ananiadou, S., S. Pyysalo, J. Tsujii, and D. B. Kell (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology* 28 (7), 381-390.

Ananiadou, S., P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell (2014). Event-based text mining for biology and functional genomics. *Briefings in functional genomics*. doi:10.1093/bfgp/elu015.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* 25 (1), 25-29.

Blaylock, N., de Beaumont, W., Allen, J., & Jung, H. (2011). Towards an OWL-based framework for extracting information from clinical texts. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp. 636-640. ACM.

Cimiano, P., U. Reyle, and J. Šarić (2005). Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering* 55 (1), 59-83.

Degtyarenko, K., P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research* 36 (suppl 1), D344-D350.

Demir, E., Cary, M. P., Paley, S., et al. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935-942.

Diehl, A. D., A. D. D. Augustine, J. A. Blake, L. G. Cowell, E. S. Gold, T. A. Gondré-Lewis, A. M. M. Masci, T. F. Meehan, P. A. Morel, A. Nijnik, B. Peters, B. Pulendran, R. H. Scheuermann, Q. A. Yao, M. S. Zand, and C. J. Mungall (2011). Hematopoietic cell types: prototype for a revised cell ontology. *Journal of biomedical informatics* 44 (1), 75-79.

Fellbaum, S. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finkel, J., T. Grenager and C. Manning (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta (2014). Pfam: the protein families database. *Nucleic Acids Research* 42 (D1), D222-D230.

Gaudet, P., P.-A. Michel, M. Zahn-Zabal, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, A. Gleizes, M. Pereira, D. Teixeira, Y. Zhang, L. Lane, and A. Bairoch (2015). The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Research* 43 (D1), D764-D770.

Gazdar, G., E. H. Klein, G. K. Pullum; I. A. Sag (1985). *Generalized Phrase Structure Grammar.* Oxford: Blackwell, and Cambridge, MA: Harvard University Press.

Godin-Heymann, N., Y. Wang, E. Slee and X. Lu (2013). Phosphorylation of ASPP2 by RAS/MAPK Pathway Is Critical for Its Full Pro-Apoptotic Function. *PLoS ONE* 8(12): e82022.

Gray, K. A., B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford (2015). Genenames.org: the HGNC resources in 2015. *Nucleic acids research* 43 (Database issue).

Gremse, M., A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg (2011). The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research* 39 (Database issue).

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. G. N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, and R. Apweiler (2004). The HUPO PSI's molecular interaction format — a community standard for the representation of protein interaction data. *Nat Biotech* 22 (2), 177-183.

Jackendoff, R. (1977). *X-bar-Syntax: A Study of Phrase Structure.* Linguistic Inquiry Monograph 2. Cambridge, MA: MIT Press.

Kim, J. D., T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii (2009). Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (BioNLP '09), pp. 1-9. ACL.

Kim, J. D., Y. Wang, T. Takagi, and A. Yonezawa (2011). Overview of Genia event task in BioNLP shared task 2011. In *Proceedings of the BioNLP*

Shared Task 2011 Workshop, BioNLP Shared Task '11, pp. 7-15. ACL.

Kim, J.-D., Y. Wang, and Y. Yasunori (2013). The Genia event extraction shared task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, pp. 8-15. ACL.

Klein, D. and C. D. Manning (2003). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems* 15 (NIPS 2002), Cambridge, MA: MIT Press.

Krallinger, M., F. Leitner, C. Rodriguez-Penagos, and A. Valencia (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome biology* 9 (Suppl 2): S4.

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88 (3), 265-266.

Manshadi, M. H., J. F. Allen, et al. (2008). Towards a Universal Underspecified Semantic Representation. 13th Conf. on Formal Grammar. Hamburg, Germany.

McCray, A. T., S. Srinivasan, and A. C. Browne (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, National Library of Medicine, Bethesda, Maryland., pp. 235-239.

Miyao, Y., K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii (2009). Evaluating contributions of natural language parsers to protein–protein interaction extraction. *Bioinformatics* 25 (3), 394-400.

Ohta, T., S. Pyysalo, and J. Tsujii (2011). Overview of the Wpigenetics and Post-translational Modifications (EPI) task of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, Portland, Oregon, USA, pp. 16-25. ACL.

Pollard, C., and I. A. Sag (1987). *Information-Based Syntax and Semantics. Volume 1. Fundamentals.* CLSI Lecture Notes 13.

Pollard, C., and I. A. Sag (1994). *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press.

Selventa (2011). *Biological Expression Lanaguge V1.0 Language Overview*, Cambridge MA 02140.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25 (11), 1251-1255.

Toutanova, K. and C. D. Manning (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000).*

The UniProt Consortium (2014). UniProt: a hub for protein information. *Nucleic Acids Research* 43 (D1), D204-D212.