MACHINE READING FOR SCIENTIFIC DISCOVERY

by

Gustave Von Hahn-Powell

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Gustave Von Hahn-Powell, titled Machine Reading for Scientific Discovery and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_________________________ Date: July 27, 2018
Sandiway Fong

_________________________ Date: July 27, 2018
Mihai Surdeanu

_________________________ Date: July 27, 2018
Clayton Morrison

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_________________________ Date: July 27, 2018
Dissertation Director: Sandiway Fong

_________________________ Date: July 27, 2018
Dissertation Director: Mihai Surdeanu

STATEMENT BY AUTHOR

SIGNED: Gustave Von Hahn-Powell

Acknowledgements

In many ways this dissertation has been a long journey of discovery—one that would certainly not have been possible without the constant support of friends, family, faculty, and peers. I came to the University of Arizona from Japan intending only to complete an MS in Human Language Technology. Meeting brilliant faculty like Sandiway Fong, Clay Morrison, Mike Hammond, Diana Archangeli, and Andy Wedel inspired me to continue my studies. I would like thank my department for its open mindedness in allowing me to pursue interdisciplinary work and for its deep commitment to the general success and welfare of all of its students.

I owe special thanks to my committee, which was jointly chaired by Sandiway Fong and Mihai Surdeanu, and rounded out by my minor advisor, Clay Morrison. Together they contribute expertise from three separate departments. Their patience and guiding questions helped shape and refine this work.

I am particularly indebted to Mihai Surdeanu who opened up a world of ideas to me. His attention and incredible work ethic challenged me to grow both personally and academically.

Thanks to Robert Poole and Zac Chapman for their hospitality when I first arrived in Tucson nervous and unprepared to start a new life. I am deeply grateful to Diana Archangeli, who for some inexplicable reason took a chance on a very naive student, and provided me with an invaluable opportunity to participate in funded research at the very beginning of my time at the University of Arizona.

In my first years as a PhD student, I was lucky to be mentored by Stacey Oberly who always made me feel welcome and unselfishly shared her time to help. Without the support of peers like Becky Sharp, Sam Johnston, and Kristen Howe, I would not have survived that formative experience under our uncompromising teacher Erwin Chan.

Collaborating with Marco Valenzuela and joining the CLU Lab changed the nature of my research. I cherish the friendship and mutual trust we fostered working countless hours alongside Dane Bell and Enrique Noriega on demanding multi-year projects.

This work was enriched by conversations with Lyn Powell, David Price, and other members of the Bill and Melinda Gates Foundation's HBGDki community who graciously volunteered their time to test the platform described in Chapter 4. Their encouragement was a source of motivation I could always draw upon, and their detailed feedback was instrumental in shaping the development of influence search.

Above all, though, it was the love and support of my wife and family that carried me through personal tragedy and the hardest moments in this journey. Were it not for these people, I would probably have left for an ashram in the Himalayas years ago.

Financial support

The work comprising this dissertation was graciously funded by the DARPA Big Mechanism program under ARO contract W911NF-14-1-0395 and by the Bill and Melinda Gates Foundation HBGDki Initiative.

Prior work

Several sections from various chapters have appeared separately as co-authored papers.

- Chapter 2 describes a multifaceted machine reading system developed in conjunction with Marco Valenzuela, Dane Bell, Enrique Noriega, Tom Hicks, Clay Morrison and Mihai Surdeanu which we first described in Valenzuela-Escárcega et al. (2017). Its application to generating biological processes relies on work from Özgün Babur and Emek Demir using the Mutex algorithm (Babur et al., 2015). The accuracy of those hypotheses in reflecting statements from the literature was evaluated by Xia Wang.
- Dane Bell deserves recognition for his tireless and careful annotation work which was instrumental in creating the causal precedence corpus described in Chapter 3. An earlier version of this corpus and assembly system was first introduced in **?**.
- The mixed domain machine reading system discussed in Chapter 4 was jointly developed with Marco Valenzuela and Mihai Surdeanu. We first presented this approach in Hahn-Powell et al. (2017).
- Finally, the short-term opportunity discovery described in the beginning of Chapter 6 is joint work with Fan Luo, Marco Valenzuela, and Mihai Surdeanu that was first presented in Luo et al. (2018) and builds upon the knowledge graph of Hahn-Powell et al. (2017).

Dedication

To the memory of my father.

Contents

List of Figures . . . 9

List of Tables . . . 12

Abstract . . . 15

Chapter 1 Introduction . . . 16
1.1 Undiscovered public knowledge . . . 16
1.2 Machine reading for scientific discovery . . . 17
1.3 Limitations of existing machine reading systems . . . 18
1.4 Key contributions . . . 19

Chapter 2 Literature-based discovery in a domain-specific setting . . . 21
2.1 Overview . . . 21
2.2 Introduction . . . 21
2.3 Related work . . . 23
2.4 Odin . . . 25
2.5 Machine reading approach . . . 29
2.6 Intrinsic evaluation: machine reading performance . . . 37
2.7 Extrinsic evaluation: discovery of biological hypotheses . . . 41
2.8 Conclusions and limitations . . . 44

Chapter 3 Causal precedence of events . . . 46
3.1 Overview . . . 46
3.2 Introduction . . . 46
3.3 Related work . . . 48
3.4 A corpus for causal precedence in the biomedical domain . . . 48
3.5 Models of causal precedence . . . 51
3.6 Results . . . 57
3.7 Discussion . . . 58
3.8 Conclusions and limitations . . . 60

Chapter 4 A platform for multi-domain scientific discovery . . . 61
4.1 Overview . . . 61
4.2 Introduction . . . 61

4.3 Related work . . . 62
4.4 Approach . . . 64
4.5 Influence graph search engine . . . 67
4.6 Evaluation . . . 70
4.7 Conclusions and limitations . . . 71

Chapter 5 Learning extraction patterns from few annotations . . . 73
5.1 Introduction . . . 73
5.2 Rule-based event extraction . . . 73
5.3 Unsupervised rule transfer . . . 75
5.4 The task . . . 76
5.5 Results . . . 79
5.6 Explaining the limitations of cross-event transfer . . . 84
5.7 Low resource settings . . . 86
5.8 Conclusions and limitations . . . 87

Chapter 6 Opportunity discovery . . . 88
6.1 Overview . . . 88
6.2 Predicting future discoveries . . . 88
6.3 Bridging distant research communities . . . 94
6.4 Conclusions and limitations . . . 103

Chapter 7 Conclusions . . . 104
7.1 Summary . . . 104
7.2 Future work . . . 104

Appendix A Reach . . . 107
A.1 Assembly annotation schema . . . 107
A.2 Example assembly output . . . 109

References . . . 111

List of Figures

1.1 According to data indexed by PubMed, biomedical publications remained fewer than 100,000 per year until 1951. The annual rate began to exceed 1,000,000 in 2011. 16
1.2 An early example of literature-based discovery from Swanson (1986b). 17

2.1 **Preprocessed sentence**. The standard preprocessing pipeline used for Odin input text produces part-of-speech tags, named entities, lemma forms, and syntactic analysis. 26
2.2 **Example Odin rules.** A simple example of an Odin grammar capturing one biochemical entity type (`Protein`) and two interactions (**Ubiquitination** and **NegativeRegulation**). This grammar also includes a taxonomy that models allowable entity and event types in this domain. 28
2.3 **Odin output mentions.** The result of applying the grammar shown in Figure 2.4 on the text in Figure 2.4. Labels in the transparent boxes refer to Odin output mentions. Nested mentions, denoted here with the label `NEGREG`, have arguments (e.g., `controlled` or `cause`) which are represented as arrows pointing to other mentions. Note that the expected type of the `controlled` argument of the **negreg** rule in Figure 2.4 is specified to be an `Event`. The **Ubiquitination** satisfies this semantic restriction. 28
2.4 Architecture of the Reach system together with a walk-through example. 29
2.5 Taxonomy of the entities and events recognized by Reach. Though abbreviated, the `Removal` events mirror those listed under `Addition`. 31
2.6 Synonymous entities are linked by grounding to a KB resource such as Uniprot (Magrane and Consortium, 2011). 32
2.7 An example of a rule template for a syntactic pattern that assumes a nominal predicate. 34
2.8 Machine reading results in the Big Mechanism evaluation. 38
2.9 The Reach output is about 12 times larger than the size of Pathway Commons. 41
2.10 Reach allows Mutex to detect 7 new candidate "driver" genes for breast cancer which are not detected otherwise, when using Pathway Commons alone, or without using any network. We observed similar results for 6 other cancers in the TCGA dataset. 42
2.11 Mutex groups for TCGA breast cancer. This graph shows the interactions of the genes in each Mutex group and their targets. The highlighted relations exist in Reach data but not in Pathway Commons. Highlighted genes are not detectable without using Reach data. 43
2.12 Mutex + Reach analysis of TCGA. 43

3.1 The distribution of assembly relation labels both within and across sentences. 50
3.2 The distribution of event pairs involving coreference across assembly relations. 51
3.3 Browser-based tool for annotating assembly relations in text. 51

3.4 Generation procedure for the cross-sentence syntactic path feature. 55
3.5 Architecture for the basic latent feature model using the minimal sentential span encompassing events 1 and 2 as input. 57
3.6 Modified architecture for a latent feature model with three-pronged input: the text of event 1 (left), the minimal sentential span encompassing events 1 and 2 (middle), and the text of event 2 (right). 57

4.1 An example of Infobox completion for the Wikipedia page on Itami Juzo. Spans of text from the accompanying article which could be used to populate the infobox are highlighted in a color that corresponds to its target field. 63
4.2 Influence relation fragments are extracted from 90,000 scientific publications within the Open Access subset of PubMed. 65
4.3 Concept mentions are simplified and linked by retaining only the lemma forms of the essential tokens (nouns, adjectives, and verbs). 66
4.4 Rules for influence events are adapted from Reach grammars pertaining to Activations and Regulations. 66
4.5 The search engine UI for the conceptual influence graph. 68
4.6 The head finding procedure used to generate a subset of the relevance scores. 69
4.7 An incomplete depiction of *Akt*'s effect on *p38*, which demonstrates multiple pathways. 70
4.8 *hypothyroidism*'s indirect role in *neuroinflammation*, which demonstrates multiple pathways with varying degrees of confidence. The dotted edges indicate that the majority of the relation's evidence is hedged (ex. *it is suggested that ...*). 70

5.1 A case of centering embedding. The abbreviated set of relations shown in this parse are Stanford collapsed dependencies (de Marneffe and Manning, 2008). `rcmod` refers to a relative clause modifier. 74
5.2 A rule finding compound noun NPs as a regular expression over one or more of part-of-speech tags where each tag must begin with `N`. 75
5.3 A pattern describing a traversal over a syntactic dependency graph. The pattern includes a repeating sequence that ultimately leads to the deepest (lowest) embedded clause. 75
5.4 Though both sentences are syntactically identical with predicates matching in valency and grammatical features, the thematic role assigned to **the book** differs (THEME in sentence above and PATIENT in the sentence below), due to the lexical semantics of the two predicates. 77
5.5 A sentence containing a **Phosphorylation** event. The event trigger, *phosphorylation*, is a nominal predicate connected to the event THEME, **CREB**, through an outgoing `prep_of` (preposition *of*) syntactic dependency relation. 78
5.6 A rule extracted from an event annotation in the training data. The rule describes a **Phosphorylation** event with a THEME as the head of an of-PP complement attached to the nominal predicate *phosphorylation*. 78
5.7 The pattern in the previous example is adapted to capture any event, given the event's label and the pattern describing that event's nominal predicates. Here `${nominalPredicate}` refers to a variable that expects a disjunction of nominal predicates. 78
5.8 The pattern in the previous example is adapted to capture a **Gene Expression** event. 79

5.9 A simple sentence with a nominal subject selected to illustrate varying degrees of lexicalization possible in a relational dependency path. 79
5.10 A lexicalized path from "ate" $\rightarrow$ "Ayu" in the sentence shown in Figure 5.9. 79
5.11 A unlexicalized path from "ate" $\rightarrow$ "Ayu" in the sentence shown in Figure 5.9. 79
5.12 A pattern with 57% precision that is well-formed, though unexpected. 82
5.13 Visualization of syntactic similarity of events. 85
5.14 Results of the experimental approach in low resource settings on the development data. 87

6.1 Procedure for temporal partitioning of the influence graph to predict future discoveries. 90
6.2 An example citation network illustrating how the outgoing citations of a venue (represented here as Journal_1) were counted. 96
6.3 Equation for truncated Singular Value Decomposition. 96
6.4 A portion of a GMap (Gansner et al., 2010) visualization depicting clusters of venues related to urology and dermatology. 97
6.5 Another portion of the same GMap visualization, this time centered around clusters related to rheumatology, neuroimaging, and microbiology. Grey lines indicate closely related venues, while clusters (or territories) demarcate highly similar groups of venues. 98
6.6 Swanson's (1986a) pioneering example of literature-based discovery in which he hypothesized that dietary fish oil could be used to treat Reynaud's disease based on their mutual relation to blood viscosity. 99
6.7 A summary of the alignment of venues to influence relations. r_1 and r_2 are relations in the influence graph linking concepts X, Y, and Z. Both r_1 and r_2 are supported by evidence (E_n) mined from the literature. Each piece of evidence originates from a publication disseminated in some venue (V_n). Together, these venues form a community (C_n) or set of communities. 100

List of Tables

2.1 Knowledge bases used by the rule-based named entity recognizer, as well as for grounding. . . 33
2.2 Common syntactic variations shared among event types. Combinations of these syntactic variations are also considered. For example, an appositive subject relative plus passivization: "Pde2, which has been found to hydrolyze Ras, activates MEK." 34
2.3 Number of rule templates in Reach's grammars. 36
2.4 Correctness of the hypotheses generated by Mutex + Reach. The "With direction" column considers strict, directional hypotheses, e.g., *GATA3 activates PTEN*. The "Ignoring direction" column considers non-directional hypotheses, e.g., either *GATA3 activates PTEN* or *PTEN activates GATA3*. . 42

3.1 The seven inter-event relation labels annotated in the corpus. The "precedes" labels are causal. Subsumption is captured with the "specifies" labels. 48
3.2 Few rules defined each deterministic model of precedence compared with the number of features for the machine learning models. 52
3.3 An overview of the four classes of feature used in the feature-based classifier. 54
3.4 Results of all proposed causal models, using stratified 10-fold cross-validation. 58
3.5 Tense and aspect for events containing verbs in the present study. 59

4.1 A summary of the relevance score metrics calculated for each deduplicated edge 69

5.1 Frequency of various **Binding** arguments. 77
5.2 Distribution of THEME instances across each event type in the training data (excluding regulations) 77
5.3 Rules extracted from the training data were applied to the corresponding event in the development set. 80
5.4 Performance of our rule transfer system on the development set of the BioNLP 2009 Shared Task. 81
5.5 Counts of the lemmatized predicate (trigger) patterns extracted in the training data. 81
5.6 Results of our top-performing experimental system on the development set of the BioNLP 2009 Shared Task. 82
5.7 Summary of rule reduction effects of precision-based thresholding. 83
5.8 Additional argument patterns through template inclusion. Precision-based thresholding of templates is a conservative effort at increasing recall. 83
5.9 The top-performing system for the task in 2009 (University of Turku). These results reflect performance on the **test** data. 84
5.10 Median scores for systems performance on the **test** data. 84
5.11 Themes across events show a low similarity in terms of their syntactic context. 84

5.12 Precision of some of the most common syntactic paths of length 1. 85
5.13 The top three most common paths for each event with a frequency ≥ 3 and precision $> 40\%$. The overall low precision of Localization and Transcription paths suggests these events have ambiguous triggers and would benefit from greater lexicalization in the representation of their paths. Note also the appearance of `_negcc`, which illustrates that the task seems to consider mentions of the form "... **expression** *of* ... *nor* **...**" to be a positive instance of an event mention, due to its implied existence in another context. 86

6.1 The total number of positive and negative examples for different values of the time threshold t. 89
6.2 60-20-20 split of positive and negative examples in the training, development, and testing partitions. 90
6.3 List of features used by the link prediction models to classify the candidate link $A \rightarrow C$, given an intermediate node B. 91
6.4 Unranked scores (precision, recall, and F1), ranked precision (P@10), and mean average precision (MAP), of several classifiers' predictions of future influence relations using backtesting at time $t = 2012$. 92
6.5 Results of an ablation experiment where one feature was removed at a time from the full AdaBoost model. 93
6.6 The top 10 influence relations predicted by our model, sorted in descending order of their informativeness score. 94
6.7 Influence pathway linking acquired mitochondrial dysfunction to Parkinson's through their mutual relationship to oxidative stress. 100
6.8 Influence pathway linking Akt phosphorylation to Parkinson's through their relationship to angiogenesis. 100
6.9 Influence pathway describing a drug interaction where one antibiotic (rifabutin) used to treat tuberculosis can interfere with the efficacy of another (clarithromycin), which is used to treat pneumonia. 101
6.10 The presence of fungal spores in the blood may increase risk of developing pneumonia. 101
6.11 Beta-catenin translocation may affect Autism by means of the Wnt pathway. 101
6.12 A protein produced by the Hepatitis virus may have a connection to Autism through its effect on the Wnt pathway. 101
6.13 Influence pathway highlighting growing evidence of a link between vitamin D and Autism. Reportedly, vitamin D supplementation can reduce the severity of Autism symptoms. 101
6.14 Through their effect on NF-kappaB, bilberries and nettle leaf extract may have some potential to treat or prevent bacterial meningitis. 101
6.15 Certain fungi produce lovastatin, which can decrease neuroinflammation. 102
6.16 CBD, one of the non-psychoactive compounds found in cannabis, may serve as a preventative supplement for kidney disorders through its effect on NF-kappaB. 102
6.17 THC, the primary pychoactive compound in cannabis, may have a therapeutic effect on treating ocular hypertension by temporarily alleviating inter ocular pressure (IOP), though the same effect does not hold true for the nonpyschoactive compound Cannabidiol (CBD). 102

6.18 Sleep fragmentation can decrease insulin sensitivity and thus increase the risk of developing diabetes. 102
6.19 Niacin, which is sometimes recommended to help manage cholesterol, can increase insulin resistance and therein raises one's risk of developing diabetes. 102
6.20 A drug used to treat schizophrenia, clozapine, has an effect on insulin signaling, which increases the risk of developing both cardiovascular disease and diabetes. 102

A.1 A description of how each column's data should be encoded for assembly annotations. 108
A.2 An annotation example for "D activates A. A inhibits the phosphorylation of B by C. B is also phosphorylated by E. Then, B is ubiquitinated by C. Then, X is ubiquitinated by Y. Ubiquitinated X binds to phoshorylated B." In addition to the fields listed here, annotators also assigned confidence scores on a 3-point likert scale (high, medium, low) to record cases of textual ambiguity. Confidence was assumed to be high if the field was left unmarked. 109
A.3 Example assembly output for two sentences taken from PMC1242134: *Akt interacts with Raf and phosphorylates it at Ser259 . Furthermore , phosphorylation of Raf by Akt inhibits activation of the Raf-MEK-ERK signaling pathway and has been shown to alter the cellular response in a human breast cancer cell line from cell cycle arrest to proliferation .* 110

Abstract

The aim of this work is to accelerate scientific discovery by advancing machine reading approaches designed to extract claims and assertions made in the literature, assemble these statements into cohesive models, and generate novel hypotheses that synthesize findings from isolated research communities.

Over 1 million new publications are added to the biomedical literature each year. This poses a serious challenge to researchers needing to understand the state of the field. It is effectively impossible for an individual to summarize the larger body of work or even remain abreast of research findings directly relevant to a subtopic. As the boundaries between disciplines continue to blur, the question of what to read grows more complicated. Researchers must inevitably turn to machine reading techniques to summarize findings, detect contradictions, and illuminate the inner workings of complex systems.

Machine reading is a research program in artificial intelligence centered on teaching computers to read and comprehend natural language text. Through large-scale machine reading of the scientific literature, we can greatly advance our understanding of the natural world. Despite remarkable progress (Gunning et al., 2010; Berant et al., 2014; Cohen, 2015), current machine reading systems face two major obstacles which impede wider adoption:

Assembly The majority of machine reading systems extract disconnected findings from the literature (Berant et al., 2014). In areas of study such as biology, which involve large mechanistic systems with many interdependent components, it is essential that the insights scattered across the literature be contextualized and carefully integrated. The single greatest challenge facing machine reading is in learning to piece together this intricate puzzle to form coherent models and mitigate information overload. In this work, I will demonstrate how disparate biomolecular statements mined from text can be causally ordered into chains of reactions (**?**) that extend our understanding of mechanistic biology. Then, moving beyond a single domain, we will see how machine-read fragments (influence relations) drawn from a multitude of disciplines can be assembled into models of children's heath.

Hypothesis generation and "undiscovered public knowledge" (Swanson, 1986a) Without a notion of research communities and their interaction, machine reading systems struggle to identify knowledge gaps and key ideas capable of bridging disciplines and fostering the kind of collaboration that accelerates scientific progress. With this aim in mind, I introduce a procedure for detecting research communities using a large citation network and derive semantic representations that encode a measure of the flow of information between these groups. Finally, I leverage these representations to uncover influence relation pathways which connect otherwise isolated communities.

1 Introduction

So many books, so little time.

Frank Zappa

The aim of this work is to accelerate scientific discovery by advancing machine reading approaches designed to extract claims and assertions made in the literature, assemble these statements into coherent models, and generate novel hypotheses that synthesize findings from isolated research communities.

1.1 Undiscovered public knowledge

Since at least 1990, there has been exponential growth in the number of academic papers published annually in the biomedical domain (Pautasso, 2012). The number of English language biomedical citations indexed by PubMed[1] alone since 1900 have now surpassed 29 million[2]. Even more staggering, over 19 million of this total was published in the period of 1990-2017.[3]

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

[2] https://www.ncbi.nlm.nih.gov/pubmed?term=(%221990%2F01%2F01%22%5BDate%20-%20Publication%5D%20%3A%20%223000%22%5BDate%20-%20Publication%5D)

[3] Oosterhaven (2015) suggests this may be due to the proliferation of journals and demonstrates statistically that any manuscript will ultimately find its way to a willing venue. At even a 30% acceptance rate, a manuscript has a 83% chance of being accepted within five rounds of resubmission.

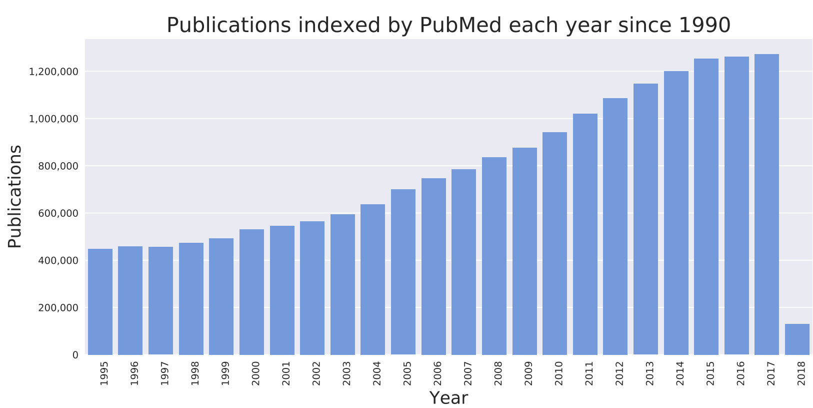


Figure 1.1: According to data indexed by PubMed, biomedical publications remained fewer than 100,000 per year until 1951. The annual rate began to exceed 1,000,000 in 2011.

Publications in the domain now exceed 1 million documents per year. This poses a serious challenge to researchers needing to understand the state of the field. It is effectively impossible for an individual to summarize the larger body of work or even remain abreast of research findings directly relevant to a subtopic. Moreover, how do all of these experimental results interact, and what broader biological context do they define? What contradictions exist within the literature? These fragmented findings, when assembled, may lead to revolutionary treatments to diseases, or obviate costly experiments in cases where the outcome may have already been convincingly established.[4]

[4] Negative and null results are rarely published or disseminated, despite their elucidatory value. Scientific progress is built upon such results. Negative findings should not be assumed to reflect deficits in methodology, design, or rigor. In fact, they may reveal methodological flaws in positive results. Wide access to negative and null results in a format amenable to machine reading, for example, could inform decisions around investment of resources, and help to avoid work that is unintentionally duplicative. When placed in a larger context, negative results have the potential to reveal key drivers in complex systems. See Gould (1993) for a more in-depth discussion of this topic, especially concerning negative findings reexamined in light of new theories.

As the boundaries between disciplines continue to blur, the question of what to read grows more complicated. Much like a puzzle, the missing piece is rarely in plain sight (see Figure 1.1). Researchers must inevitably turn to information extraction techniques to summarize findings, detect contradictions, and illuminate the inner workings of complex systems. The need for machine reading and, crucially, automatic methods for integrating reported findings has never been greater. Indeed, it is the only path forward.

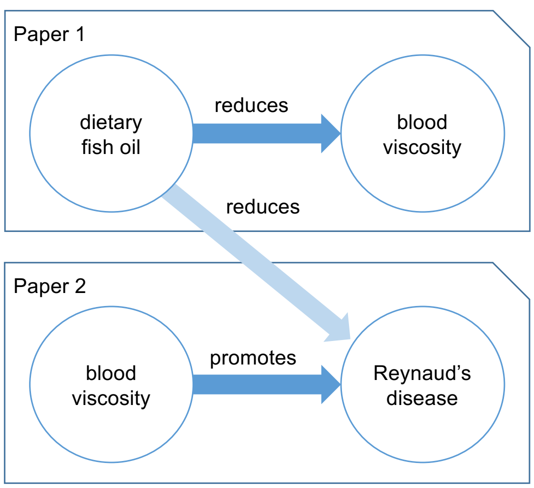


Figure 1.2: Swanson's (1986b) pioneering example of literature-based discovery in which he hypothesized that dietary fish oil could be used to treat Reynaud's disease based on their mutual relation to blood viscosity. The key observations serving as motivation for the proposed treatment were found in the titles of seemingly unrelated papers. Swanson's early discoveries were made through manual analysis; however, the body of scientific literature has grown by an order of magnitude since that time.

1.2 *Machine reading for scientific discovery*

Machine reading is a research program in artificial intelligence centered on teaching computers to read and comprehend natural language text. Swanson (1986a,b) pioneered this work in the 1980s by demonstrating how the titles of publications contained isolated discoveries, that, when linked, revealed novel treatments for diseases and hypothesized underlying causes. Through large-scale machine reading of the scientific literature, we can greatly advance our understanding of the natural world and construct models that reflect human reasoning. While such work is far from over, great strides have already been made in areas such as question answering (Gunning et al., 2010), micro-reading (Berant et al., 2014), and systems biology [5].

[5] Cohen, P. R. (2015). DARPA's Big Mechanism program. *Physical biology*, **12**(4), p. 045008

1.3 *Limitations of existing machine reading systems*

Despite remarkable progress, current machine reading systems have a number of limitations and deficits. Chief among these is the assembly of disconnected facts mined from the literature, a process which consolidates information and enables novel discoveries through the integration of previous findings.

Assembly The majority of machine reading systems extract isolated facts or statements from the literature (Berant et al., 2014). In areas of study such as biology, which involve large mechanistic systems with many interdependent components, it is essential that the insights scattered across the literature be contextualized and carefully integrated. The single greatest challenge facing machine reading is in learning to piece together this intricate puzzle to form coherent models and mitigate information overload.

Hypothesis generation and "undiscovered public knowledge" (Swanson, 1986a) Without a notion of research communities and their interaction, machine reading systems struggle to identify knowledge gaps and key ideas capable of bridging disciplines and fostering the kind of collaboration that accelerates scientific progress.

Predictive power Comprehensive retrieval of historical discoveries across the published literature is the initial goal of machine reading, but what can such systems tell us about discoveries to come or fruitful avenues of critical research? Ultimately, we will expect these systems to generate novel hypotheses, highlight methodological flaws, and aid researchers in prioritizing and designing experiments.

Domain specificity While a number of systems have focused on domain independent information extraction, the dialogue in individual scientific disciplines is nuanced and often requires special treatment (Bell et al., 2016). In many cases, training data for information extraction may be sparse or nonexistent. This problem is further complicated by new and emerging fields which straddle multiple domains.

Interpretability and extensibility Subject matter experts must be able to interact with machine reading systems. Not only must the machine's reasoning be comprehensible to the human expert, the system itself should ideally be extensible (e.g., capable of or readily adapted to fulfilling new information needs).

1.4 Key contributions

This thesis focuses on assembling claims mined from scientific literature and synthesizing these findings to generate novel hypotheses.

Assembly Moving beyond isolated statements, I explore challenges in assembling the output of machine reading, before presenting a corpus of causal precedence, and demonstrating how fragmented biomolecular statements mined from text can be causally ordered to form meaningful chains of reactions that deepen our understanding of mechanistic biology. This work is covered in **Chapter 3**. Next, in **Chapter 4**, I show how a machine reading system developed for the biomolecular domain can be adapted to simultaneously cover multiple domains where claims mined from a disparate body of literature are automatically assembled to form a conceptual *influence graph*[6]. Built upon this knowledge graph is a platform for scientific discovery where researchers can rapidly and collaboratively construct models of complex public health issues through a machine-aided review of the literature.

[6] A graph consisting of statements which describe or suggest a causal effect, or what Hearst (1999) calls "chains of causal implication." Influence graphs define the causal landscape surrounding topics such as children's health (Hahn-Powell et al., 2017), soil quality, and cancer biology (Cohen, 2015).

Hypothesis generation and "undiscovered public knowledge" I demonstrate how semantic representations of research communities can be derived from a large citation graph and applied to an influence graph to reveal chains of findings which bridge otherwise isolated communities. Described in **Chapter 6**, this community-aware approach to hypothesis generation is accompanied by a set of diverse examples.

Predictive power To aid in harvesting low-hanging opportunities for validating research, the first section of **Chapter 6** describes joint work with colleagues on a method for predicting near-term future discoveries based on trends observed in a large influence graph.

Domain specificity Highlighting the explanatory power of specialized machine reading approaches, **Chapter 2** describes a high precision machine reading system tailored to the biomolecular domain that has been used and extended by biomedical researchers to reveal plausible targets for treating several cancers (Valenzuela-Escárcega et al., 2017).

Interpretability and extensibility In an effort to address the challenges of interpretability and extensibility, **Chapter 5** investigates how the hand-crafted rule-based grammars which drive the machine reading systems described in Chapters 2 and 4 can be inferred from a relatively small number of annotations, generalized to templates, and finally selectively applied in a performance-aware fashion to recognize other event types.

2 Literature-based discovery in a domain-specific setting

2.1 Overview

This chapter introduces Reach(REading and Assembling Contextual and Holistic mechanisms from text), a high precision machine reading system tailored to the biomolecular domain first introduced in Valenzuela-Escárcega et al. (2017) [1], capable of recognizing textual mentions of a wide variety of biomolecular events. Reach performs domain-specific coreference resolution (Bell et al., 2016) and detects both hedging and negation for every statement it extracts from text. Our system is built on Odin [2], a domain-independent framework for information extraction which allows for the specification of rules over syntactic structures, as well as sequences of contiguous or noncontiguous tokens and their linguistic attributes. Odin rules operate in a grammar cascade and can incorporate custom taxonomies, as well as support hierarchical semantic constraints on arguments to some event predicate. Here we describe not only the design of Reach, but also how biomedical researchers used the system in conjunction with patient data to reveal plausible targets for treating several cancers.

[1] Valenzuela-Escárcega, M. A., Ö. Babur, G. Hahn-Powell, D. Bell, T. Hicks, E. Noriega-Atala, X. Wang, M. Surdeanu, E. Demir, and C. T. Morrison (2017). Large-scale automated reading with Reach discovers new cancer driving mechanisms. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*, pp. 201–203

[2] Valenzuela-Escárcega, M. A., G. Hahn-Powell, T. Hicks, and M. Surdeanu (2015b). A Domain-independent Rule-based Framework for Event Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Software Demonstrations*, pp. 127–132. ACL-IJCNLP 2015. DOI: 10.3115/v1/P15-4022

2.2 Introduction

Each year, more than a million new articles are added to biomedical literature. This far exceeds the processing capacity of even teams of human domain experts, limiting our ability to truly understand many diseases. We present Reach, a system for automated, large-scale reading of biomedical papers that can extract mechanistic descriptions of biological processes with relatively high precision and throughput. We demonstrate that combining the extracted pathway fragments with existing biological data analysis algorithms that rely on curated models helps identify and explain a large number of previously unidentified mutually exclusive altered signaling pathways in seven different cancer types. This work shows that combining curated "big mechanisms" with extracted "big data" can lead to a causal, predictive understanding of cellular processes and unlock downstream applications.

Despite more than a decade of human curation efforts dedicated to building and extending resources such as Pathway Commons[3], we estimate only 1–3% of the literature has been covered to date, and this gap widens every day.[4] Much of the mechanistic knowledge described in the literature remains hidden in unstructured text. This poses a serious challenge to biological data analysis. As a concrete example, consider the detection of "driver" mutations in cancer. One widely recognized observation is that, given a cohort of patients, some driver alterations will exhibit a mutually exclusive pattern. That is, the number of patients that have both alterations will be smaller than what is expected by chance. This often happens because these alterations unlock the same cancer driving pathways and the positive selection of one diminishes substantially when the other is present. In other words, one is sufficient. Prior knowledge of biomolecular pathways can be used to improve the accuracy of these methods by limiting the search space and reducing the loss of statistical power due to multiple hypothesis testing correction. It also provides mechanistic explanations of the observed correlations (Babur et al., 2015). Recall, however, can be low due to the aforementioned database coverage issue.

[3] www.pathwaycommons.org

[4] Internal analysis of the Pathway Commons team.

Researchers are thus faced with a choice between high-coverage methods without priors that do not provide mechanistic explanations or low-coverage, prior-based methods that may overlook some key events.

To fully answer these complex biological questions, we propose a natural language processing (NLP) approach that captures a system-scale, mechanistic understanding of cellular processes through automated, large-scale reading of scientific literature, and demonstrate that this approach leads to the discovery of novel biological hypotheses for multiple cancers. We call our approach Reach (REading and Assembling Contextual and Holistic mechanisms from text).

Our approach has two important contributions. The first contribution is the demonstration that the combination of "big data" that is produced by machines and "big mechanisms" that were manually curated yields novel knowledge that is otherwise missed. In particular, we show that Reach can substantially improve the inference capacity of existing biological data analysis algorithms that previously relied solely on manually-curated pathway databases such as Pathway Commons (PC). Here, we extended the Pathway Commons human-curated pathways with over one million biochemical interactions extracted by Reach from all papers in the Open Access subset of PubMed (as of June 2015). Using this combined prior network we were able to identify a large number of previously unidentified, but highly statistically significant mutually exclusively altered signaling modules in TCGA cancer datasets using the Mutex algorithm (Babur et al., 2015). A manual evaluation of these modules reveals that between 65 and 80% of the pathway fragments discovered by Reach are correct, and they indeed help uncover novel biological hypotheses within the corresponding cancer context.

Our second contribution is the machine reading approach itself. The core of Reach is a cascade of automata that relies on *compact* and *interpretable* grammars that extract entities (e.g., proteins) and events (e.g., biochemical interactions) of interest. This guarantees that the reading model can be understood, modified, and extended by domain experts. This compact grammar is efficiently applied at runtime, an important benefit in our "big data" setup. On average, we process a single paper in 4.5 seconds, though our software can be easily parallelized to take advantage of cluster computing and multi-core hardware. Additionally, Reach captures complex natural language phenomena such as coreference, and can interpret event polarity in statements with nested contrasts (for example, "a *reduction* of *increased* phosphorylation led to . . . "). An independently administered evaluation found that Reach extracts cancer signaling pathways at relatively high precision and at a throughput capable of reading the whole biological literature in short order.

2.3 Related work

Reach builds upon the tremendous body of work in language technology applied to bioinformatics that was developed in the past two decades. We summarize the major trends that influenced our work below, but for a more comprehensive background we recommend reviews of the field such as [5].

[5] Cohen, K. B. and D. Demner-Fushman (2014). *Biomedical Natural Language Processing*, volume 11. John Benjamins Publishing Company. DOI: 10.1075/nlp.11

Due to the above-mentioned information explosion in biomedical research, it is imperative to develop reliable, automated methods to extract information from this literature and make it available in a structured fashion. The BioNLP shared tasks and associated workshops were organized to advance research in this area (Kim et al., 2009, 2011; Nédellec et al., 2013). Many systems have participated in this shared task, broadly representing two directions: rule-based and machine learning methods.

Rule-based information extraction systems have been successful in the biomedical domain. These systems gained noteriety with the advent of FASTUS (Appelt et al., 1993), which was implemented as a cascade of finite state automata (FSA), where each FSA captured a "layer" in the task to be addressed (e.g., entities, events), and was defined through a grammar that aggregated multiple rules. Systems such as FASTUS tend to rely on shallow linguistic structure for efficiency. Inspired by the ideas promoted by FASTUS, one of the first rule-based information extraction systems to target the biomedical domain was Blaschke et al. (1999), which focused on extracting protein-protein interactions. Devised by biologists, the system searches for mentions of proteins separated by a term known to signal their interaction. The extracted protein-protein interactions were then assembled into a small interaction graph with a high degree of accuracy.

While Blaschke et al. (1999) demonstrated the effectiveness of lexicalized patterns, deeper linguistic analysis affords certain advantages such as better generalization. Kilicoglu and Bergler (2009) used a concise set of rules over deep linguistic structure (dependency parses) to detect nine types of biochemical events. This system was one of the top performers in the BioNLP 2009 shared task on event extraction.

An important trend in information extraction is, of course, the use of machine learning. These approaches can be classified in two sub-classes: supervised learning, where the machine learns from data manually annotated by domain experts, and distant supervision, where training data is automatically generated by aligning a database of known facts, e.g., protein-protein interactions, with relevant texts, e.g., biomedical publications discussing such interactions (Mintz et al., 2009). The first approach that applied machine learning to biomedical information extraction was proposed by Craven and Kumlien (1999). Notably, this is also the first work to use distant supervision for information extraction. Björne et al. (2009) proposed a supervised machine learning approach for biomedical IE, which obtained the best results at the BioNLP 2009 shared task on event extraction. Since then, several efforts have improved upon its performance (Miwa et al., 2010; McClosky et al., 2012; Miwa et al., 2012; Bui and Sloot, 2012; Venugopal et al., 2014). Notably, the top performers at the more recent editions of the BioNLP shared task rely on machine learning (Björne and Salakoski, 2013; Miwa and Ananiadou, 2013).

Reach builds upon this previous work in several ways. First, we propose a declarative rule-based approach that is inspired by and improves upon this body of work, using a framework designed to build grammars that are concise, interpretable, and which can mix deep and shallow syntactic analysis. Our system also addresses important phenomena that have generally been ignored in previous work (e.g., coreference resolution and event polarity). Finally, the approach we present here can be combined with machine learning approaches to discover relevant grammars automatically. Our experiments indicate that such hybrid approaches can be constructed at minimal cost and are successful (Valenzuela-Escárcega et al., 2016). While others have looked at the overlap between curated models and literature-derived networks (Spranger et al., 2016; Kim et al., 2015), our work is the first to carefully examine whether the biochemical pathways extracted by the machine can be successfully combined with human-curated models in the context of a specific analytical task.

2.4 *Odin*

The NLP method described here was developed using Odin, a domain-independent information extraction framework we introduced in previous work [6]. Odin operates by applying a cascade of automata (each compiled from a different grammar) over free text to incrementally extract relevant structures such as entities (e.g., proteins in the biomedical domain) and events (e.g., biochemical interactions). One fundamental contribution of Odin is that these grammars operate over both surface text and syntactic dependency graphs. These grammars are defined using a novel declarative language that is compiled into a collection of finite state automata.

[6] Valenzuela-Escárcega, M. A., G. Hahn-Powell, T. Hicks, and M. Surdeanu (2015b). A Domain-independent Rule-based Framework for Event Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Software Demonstrations*, pp. 127–132. ACL-IJCNLP 2015. DOI: 10.3115/v1/P15-4022; and Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2016). Odin's Runes: A Rule Language for Information Extraction. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC

The minimal requirement for Odin input is text that is tokenized into distinct words. However, as shown in Figure 2.4, Odin grammars frequently use syntax to capture relevant domain structures. It is therefore common that text that is analyzed by Odin be first annotated using standard NLP tools such as text segmentation, part-of-speech (POS) tagging, named entity recognition (NER), and syntactic parsing. We provide a brief description of each of these tasks below.

Text Segmentation

Text segmentation refers to the task of dividing text into meaningful units such as sentences and words.

Part-of-speech tagging

Part-of-speech (POS) tags are categories that capture syntactic functions of words, e.g., nouns, verbs, and prepositions. POS tags are generally assigned using sequence models such as conditional random fields (CRFs).

Named-entity recognition

Named-entity recognition (NER) is another sequence modeling task that identifies sections of text referring to mentions of entities of interest. Examples of such entities are names of people, places, and organizations (in an open-domain scenario), or proteins, simple chemicals and cellular locations (in the biomedical domain).

Because named entities can be composed of multiple words, they are typically labeled using the BIO notation [7]. This representation uses the prefixes `B-` and `I-` to denote words that are at the *beginning* or *inside* an entity group. `O` is used to mark tokens that are *outside* of any group (e.g., nonentities).

[7] Ramshaw, L. and M. Marcus (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*

Syntactic parsing

Syntactic parsing captures intra-sentence grammatical relations (e.g., subject, object). There are multiple possible representations for these relations. Odin uses the Stanford typed dependencies representation [8]. Stanford dependencies consist of directed and labeled binary relations between pairs of words: a head and its dependent. For example, the relation between a verb and its subject is encoded as the nominal subject (`nsubj`) relation where the verb is the head and the subject is the dependent. Furthermore, Odin uses the collapsed dependencies representation, where dependencies involving prepositions and conjunctions are collapsed in order to get direct connections between content words. This facilitates the development of extraction rules.

[8] de Marneffe, M.-C. and C. D. Manning (2008). The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Coling 2008 Organizing Committee

An example of a sentence annotated by the above preprocessing tools is shown in Figure 2.4. The text in the figure has been tokenized into words. The colored text directly above each word indicate its POS tag. The bottom row shows named-entity tags in BIO notation. The arrows connecting the tokens indicate the directed edges of the dependency graph, connecting the head to the dependent. Note that Odin is agnostic to the implementation of the NLP preprocessing tools. We take advantage of this flexibility by using tools customized for the biomedical domain.

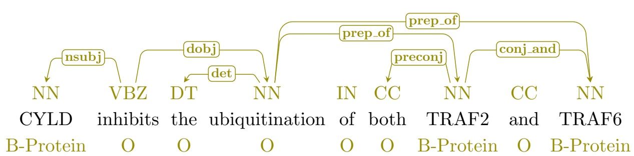


Figure 2.1: **Preprocessed sentence**. The standard preprocessing pipeline used for Odin input text produces part-of-speech tags, named entities, lemma forms, and syntactic analysis.

To highlight Odin's capabilities, we show a simple example of an Odin grammar containing three rules in Figure 2.4. At runtime, Odin applies these grammars over preprocessed text, and records all text segments where rules match. We refer to these matches as *mentions*. The example grammar shown in Figure 2.4 operates as follows:

- The `ner` rule converts the BIO output of an external NER tool into Odin entity mentions labeled `Protein`. This rule is a simple example of a surface rule[9], operating over the sequences of tokens and annotations of those tokens.

[9] In NLP terminology, a surface rule denotes a pattern over shallow linguistic features (part-of-speech tag, lemma, entity label, etc.) associated with the words in a sequence.

- The `ubiq` rule matches a **Ubiquitination** event, which is anchored around a nominal predicate (`trigger`), "ubiquitination", and has two arguments: a mandatory `theme`, which is syntactically attached to the verbal trigger through the preposition "of", and an optional `cause`, attached to the trigger through the preposition "by". Unlike entity mentions, Odin event mentions keep track of their participants (e.g., themes and causes), in addition to the matching text. The resulting event mention is assigned the **Ubiquitination** label. The hypernym labels `Simple_event` and `Event` are added automatically, according to the provided taxonomy (see Figure 2.4 for an example taxonomy). The taxonomy is a tree structure, optional in Odin, which defines a hierarchy of labels to be assigned to each mention.

- The `negreg` rule implements a negative regulation driven by a verbal predicate. Note that one of the arguments is an event produced by the `ubiq` rule. The Odin framework guarantees that the rules will be executed in the correct order, even when the order is not explicitly specified. (For details on explicitly determining the execution order we refer the reader to the Odin manual [10].)

[10] Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2015c). Description of the Odin Event Extraction Framework and Rule Language

Figure 2.4 shows the output of applying the grammar in Figure 2.4 over the sentence shown in Figure 2.4. This simple example highlights several of Odin's advantages. First, Odin easily captures nested events, such as the negative regulations in the figure, simply by requiring an event's argument to be another event (consider the `theme` in the `negreg` rule). Second, this example highlights one of the advantages of using syntax: because the `ubiq` rule in Figure 2.4 uses the `prep_of` relation to denote participants in the ubiquitination reaction, Odin automatically captures the fact that there are two ubiquitination reactions in the sentence in Figure 2.4 (due to the repeated `prep_of` dependencies), which would otherwise be cumbersome to capture using surface patterns.

Writing extraction rules over syntactic structures yields compact models that generalize well without sacrificing accuracy (Valenzuela-Escárcega et al., 2015b). However, syntax-based rules assume the availability of a reliable syntactic parser. This is not always true, especially in domain-specific settings. In such situations it is beneficial to mix syntax and surface rules, taking advantage of both the flexibility of syntax and the robustness of surface forms. Odin can freely mix syntax- and surface-based rules in the same grammar. As a sim-

```
taxonomy:
  - Entity:
    - Protein
    - SimpleChemical
  - Event:
    - SimpleEvent:
      - Phosphorylation
      - Ubiquitination
    - NestedEvent:
      - PositiveRegulation
      - NegativeRegulation

rules:
  - name: ner
    label: Protein
    type: token
    pattern: |
      [entity="B-Protein"][entity="I-Protein"]*

  - name: ubiq
    label: Ubiquitination
    pattern: |
      trigger = ubiquitination
      theme:Protein = prep_of
      cause:Protein? = prep_by

  - name: negreg
    label: NegativeRegulation
    pattern: |
      trigger = [lemma=inhibit & tag=/^V/]
      controlled:Event = dobj
      cause:Protein = nsubj
```

Figure 2.2: **Example Odin rules.** A simple example of an Odin grammar capturing one biochemical entity type (`Protein`) and two interactions (**Ubiquitination** and **NegativeRegulation**). This grammar also includes a taxonomy that models allowable entity and event types in this domain.

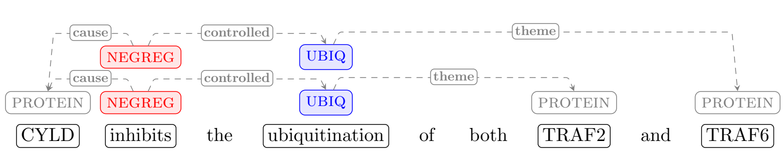


Figure 2.3: **Odin output mentions.** The result of applying the grammar shown in Figure 2.4 on the text in Figure 2.4. Labels in the transparent boxes refer to Odin output mentions. Nested mentions, denoted here with the label `NEGREG`, have arguments (e.g., `controlled` or `cause`) which are represented as arrows pointing to other mentions. Note that the expected type of the `controlled` argument of the **negreg** rule in Figure 2.4 is specified to be an `Event`. The **Ubiquitination** satisfies this semantic restriction.

ple example, the `ner` rule in Figure 2.4 uses a surface-based pattern, whereas the other rules are syntax-based.

2.5 *Machine reading approach*

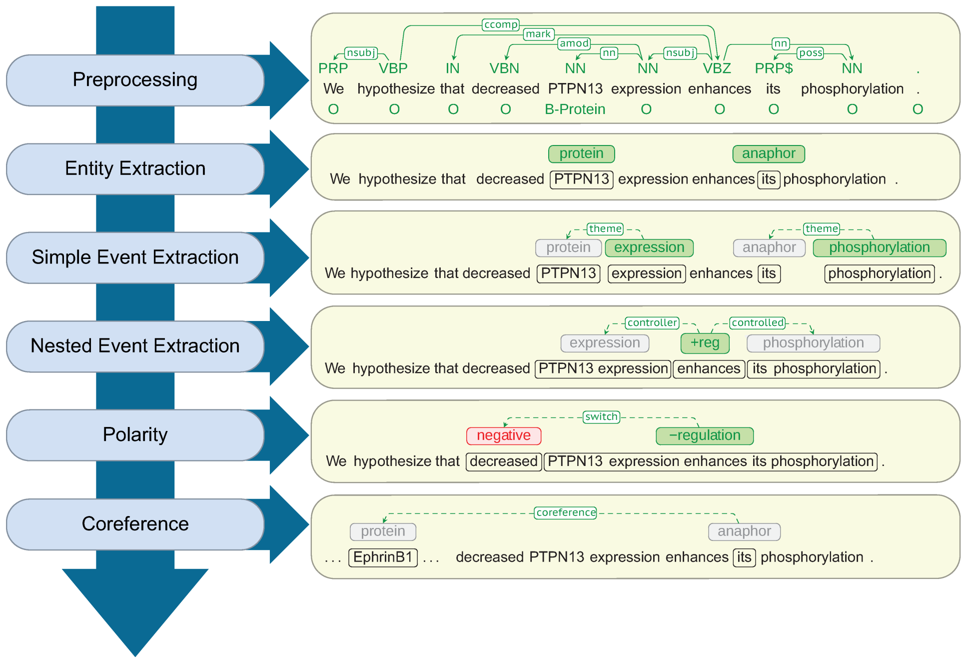


Figure 2.4: Architecture of the Reach system together with a walk-through example.

At a high level, Reach uses a cascade of rule-based and statistical techniques to read the content of a paper and identify mentions of molecular events that describe fragments of a signaling pathway. The steps of this sequence proceed from from low- to high-complexity representations, each building on the output of the previous steps, as shown in Figure 2.5.

The representation of these mentions is constructed internally in a format inspired by the BioPAX standard language (Demir et al., 2010). Notably, Reach can represent detailed biochemical conversions where entities go through "state" changes such as becoming phosphorylated, or changing their sub-cellular location. Reach also represents controllers or catalysts of these conversions, when they are mentioned in text. Similar to BioPAX, these are represented using a *composite* structure where events can have other events as their participants, allowing for arbitrarily complex logic. An important extension of BioPAX that Reach implements is the extraction of higher-level control relations between entities (e.g., "KRAS activates p53"). Although such relations are biologically ambiguous relative to a mechanistic conversion representation (e.g., the above example summarizes the biological mechanism "KRAS promotes phosphorylation of p53 on Ser37"), they provide valuable information to domain experts.

In the following sections we describe details of the Reach architecture components in Figure 2.5 that are responsible for the extraction of these mechanism fragments.

Preprocessing

Reach first preprocesses the text with NLP tools specifically modified for the biomedical domain. Preprocessing includes sentence and word segmentation, part-of-speech (POS) tagging, and syntactic parsing.

The sentence and word segmentation step detects both sentence and word boundaries in the input text. There are subtle but important differences between the tokenization of open-domain text and biomedical content. For example, dashes that occur within a word are not considered separators when segmenting open-domain text, but they tend to function as word separators in biomedical texts. For example, segmenting the text "GAP-mediated" at the dash is crucial for the downstream components to understand that this text contains a catalysis driven by GAP. Similarly, not considering the dash as separator would prohibit the downstream components from recognizing members of protein complexes, which typically appear as dash separated in text. To handle these phenomena, a custom segmenter was developed in-house, following the tokenization specification of the GENIA corpus (Kim et al., 2003).

For POS tagging and syntactic parsing, Reach uses Stanford's CoreNLP toolkit (Manning et al., 2014), which has been trained using a combination of two corpora: the Penn Treebank, a corpus that merges several non-biomedical genres such as IBM computer manuals and Wall Street Journal articles (Marcus et al., 1993; Taylor et al., 2003), and the GENIA corpus, which is a manually annotated corpus of 2000 MEDLINE abstracts (Kim et al., 2003). Including the GENIA annotated documents as part of the parser's training corpus makes the parser more robust to syntactic structures often found in biomedical literature.

Entity extraction

Next, a custom named entity recognizer (NER) component is used to recognize mentions of relevant physical entities by type, such as protein family, cellular component, simple chemical, site, and gene or gene product (this last category includes genes and proteins). The complete list of entities recognized by Reach as well as the biochemical events described later, is listed in the taxonomy in Figure 2.5.

Figure 2.5: Taxonomy of the entities and events recognized by Reach. Though abbreviated, the `Removal` events mirror those listed under `Addition`.

The custom NER uses a hybrid approach that combines a rule-based component with a statistical one. The rule-based component recognizes all mentions of known entity names (and their synonyms) from the knowledge bases (KBs) shown in Table 2.5. Additional grammar rules were written to capture entities that are not adequately covered by these KBs, such as cellular components or sites of biochemical reactions. The statistical model is implemented using CoreNLP's conditional random field (CRF) sequence classifier, trained on the BioCreative corpus (Hirschman et al., 2005). This dataset supports only mentions of gene or gene products. The hybrid NER combines the output of the two components, prioritizing the rule-based component when overlaps are detected.

Next, Reach "grounds" the biochemical entities discovered by linking the textual mentions to ids of actual entities in the knowledge bases shown in Figure 2.5 and Table 2.5. For example, the protein mention "MEK1" can be linked to the ID `Q02750` in the Uniprot knowledge base.

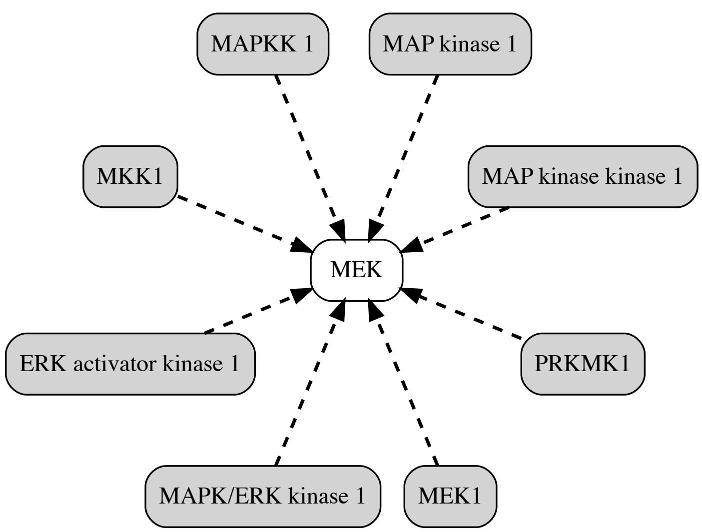


Figure 2.6: Synonymous entities are linked by grounding to a KB resource such as Uniprot (Magrane and Consortium, 2011).

Lastly, Reach detects mentions of gene mutations and protein post-translational modifications (PTMs), and attaches them to the corresponding textual mentions of these biochemical entities. This is implemented with a subsequent grammar that focuses on detecting changes of states in the previously extracted entity mentions, e.g., from the text "wild type EHR" Reach extracts the state Wild Type (i.e., non mutated) and attached it to the previously extracted entity

Entity Type	*Database*	*URL*
Protein	UniProt	`www.uniprot.org/`
Protein families	InterPro	`www.ebi.ac.uk/interpro/`
Simple Chemicals	HMDB	`www.hmdb.ca/`
Simple Chemicals	ChEBI	`www.ebi.ac.uk/chebi/`
Sites	InterPro	`www.ebi.ac.uk/interpro/`

Table 2.1: Knowledge bases used by the rule-based named entity recognizer, as well as for grounding.

mention "EHR".[11]

[11] We will extend this component in future work to include binding sites and fragments.

Event extraction

Once Reach has determined which entities are mentioned in the text, it extracts the biochemical processes in which they participate. We use a two-step bottom-up strategy for event extraction, following biochemical semantics inspired by BioPAX. First, we identify biochemical reactions that operate directly on entities, temporarily ignoring their catalysts and other controllers (e.g., phosphorylation of a protein). Following NLP terminology, we call these events "simple". Second, we find the processes that control these conversions, e.g., the catalysis of this phosphorylation by a kinase. We call these events "nested", due to the fact that they have other events as their targets (e.g., the above catalysis operates on a **phosphorylation** simple event).

One notable contribution of this work is the small number of rules used for event extraction. This is achieved by first identifying several general syntactic variations shared among event mentions, and then reusing the same syntactic structures for all event types. Table 2.5 describes 10 syntactic variations used in this work, together with examples for each.

We implement the above intuition using templates expressed in the Odin information extraction rule language (Valenzuela-Escárcega et al., 2015c, 2016). Odin templates enable expression of rules representing parameterized patterns. Figure 2.5 shows a template of the `ubiq` rule introduced in Figure 2.4. As another example, we used one template to describe the declarative syntactic pattern in Table 2.5, but left the actual verb as a parameter to be instantiated later. The particular verb to be used at runtime is initialized with specific values for the different event types (e.g., "phosphorylate" for phosphorylation events).

In all, we support 12 different types of simple events, as highlighted in Figure 2.5. Nine of these are biochemical reactions: phosphorylation, ubiquitination, hydroxylation, sumoylation, glycosylation, acetylation, farnesylation, ribosylation, and methylation. All of

```
name: ${EntityType}-nominal-predicate
label: ${EntityType}
pattern: |
  trigger = ${nominalPredicates}
  theme:Protein = prep_of
  cause:Protein? = prep_by
```

Figure 2.7: **Example Odin rule template.** An example of a rule template for a syntactic pattern that assumes a nominal predicate. Variables are denoted using `${}` (`${EntityType}` and `${nominalPredicates}` in this example). One fully instantiated form of this template is the `ubiq` rule given in Figure 2.4.

Name	*Description*	*Example*
Declarative	The theme (the thing acted on by the verb) is the direct object of a verb.	"Smurf1 and Smurf2 degrade and ubiquitinate RhoA."
Passive	The theme is the syntactic subject of a verb phrase.	"RhoA is ubiquitinated and degraded by Smurf1 and Smurf2."
Prepositional Nominalization	The trigger is in noun form and entities are in prepositional phrases.	"The ubiquitination and degradation of RhoA by Smurf1 and Smurf2 increased."
Object Nominalization	The trigger is in noun form and with the theme forms a noun-noun compound.	"RhoA ubiquitination and degradation by Smurf1 and Smurf2 increased."
Subject Nominalization	The trigger is in noun form and with the cause forms a noun-noun compound.	"Smurf1 ubiquitination and degradation of RhoA increased."
Subject Relative Clause (+ optional Apposition)	The trigger and theme are located in a relative clause which modifies the cause.	"Its many abnormal phenotypes can be rescued via Pde2, which specifically hydrolyzes cAMP."
Object Relative Clause (+ optional Apposition)	The trigger and cause are located in a relative clause which modifies the theme.	"We measured transcription activation in the presence of cAMP, which is hydrolyzed by CRP."
Subject Apposition	The cause is in an appositive phrase.	"Via yeast two-hybrid screening, we found that a novel protein, A20, binds to ABIN."
Object Apposition	The theme is in an appositive phrase.	"Via yeast two-hybrid screening, we found that A20 binds to a novel protein, ABIN"
Paraphrastic Causative	The trigger is separated from an entity by a verb.	"Smurf1 causes the degradation of RhoA."

Table 2.2: Common syntactic variations shared among event types. Combinations of these syntactic variations are also considered. For example, an appositive subject relative plus passivization: "Pde2, which has been found to hydrolyze Ras, activates MEK."

these reactions involve the covalent modification of a protein. The difference between these events and the PTMs extracted in the previous step is that these events refer to the actual act of modifying the protein by attaching a functional group to it, and the PTMs described in the previous step refer to proteins that have already been modified (potentially as a result of simple events mentioned previously in the paper).

The three remaining simple events are 1) translocation, which refers to the act of transporting an entity between two cellular locations; 2) binding, which is the process of assembling a complex from two or more proteins; and 3) hydrolysis, which is the separation of chemical bonds by the addition of water. Hydrolysis captures activities like cleavage and degradation.

Nested events are processes that control other events, such as catalysis and inhibition. Reach recognizes both positive (e.g., "promotes") and negative (e.g., "inhibits") controls. It is also possible to chain the control logic, e.g., the co-modulation of a catalysis. Following BioNLP terminology (Kim et al., 2009), we collectively call these type of events "regulations" for simplicity.

Reach also recognizes mentions of "activations", i.e., higher-level interactions that describe the direct control of an entity's activity, e.g., "A activates B", where A and B are proteins. These are structurally very similar to regulations with the exception that the "controlled" participant is an implied downstream activity of a biochemical entity. These are not supported in BioPAX by design due to the inherent semantic ambiguity: proteins can have multiple, overlapping "activities". Reach supports them because they are abstractions frequently used to summarize the result of a sequence of steps in a signaling pathway. These activations are not as useful as regulations when considered in isolation, but they provide valuable information, including the author's high-level interpretation of the discussed mechanism, and indirect dependencies between proteins. In the next section, we demonstrate how to use this information to discover latent explanations for cancer drivers.

Similar to simple events, nested events conform to the syntactic patterns shown in Table 2.5. Capitalizing on these patterns, the extraction system was implemented in Odin using 154 unique rule templates, as shown in Table 2.5.

•tType	*Syntax*	*Surface*	*Total*
Entities	0	15	15
Generic entities	0	2	2
Modifications	0	6	6
Mutants	0	9	9
Total	0	32	32
Simple events	15	11	26
Binding	30	7	37
Hydrolysis	8	2	10
Translocation	12	0	12
Positive regulation/activation	16	4	20
Negative regulation/activation	14	3	17
Total events	95	27	122
Total	95	59	154

Table 2.3: Number of rule templates in Reach's grammars.

Complex natural language phenomena

In addition of the event and entity extraction grammars described previously, Reach also recognizes complex phenomena that are difficult to detect with rules alone, namely hedging and coreference.

Hedging It is not uncommon for authors in this domain to soften claims through the use of hedging. Reach uses a series a simple rules based on modal verbs and selected adverbs to assess the definitiveness of an author's claim. Our system retains any matches to such rules to enable later analysis and statement filtering in downstream applications.

Coreference resolution Coreference, (i.e., the ability for different mentions in text to refer to the same real-world entity or event), is common in the biomedical domain. Resolving these coreference links leads to greater recall in information extraction, but it is rarely pursued in the biomedical domain. Coreference applies to both entities and events, and often reaches across sentence boundaries, as in the following examples, in which the **bold text** refers back to the *italicized text*. The correct coreference resolution in each case allows a further event to be extracted.

- "In the current study, we describe the phosphorylation, localization, and genome-wide regulatory functions of *HP1γ* in gonadal tissue, gametes, and the pre-implantation embryo. We demonstrate that phosphorylation of **this protein** at S83, which occurs in response to Aurora A, is necessary for supporting proper mitotic cell division in cells from the sperm lineage."

- "When Wnt signaling or Cdc42 activity was blocked, the induced, but not the basal level of this interaction, was lost, suggesting both Wnt and Cdc42 activities are required to promote a *Dvl2/aPKC interaction* after scratching. In contrast, aPKC inhibitors did not block **this interaction**, suggesting aPKC activity was not required for Dvl2/aPKC complex formation."

Inspired by Lee et al. (2013), we adopted an architecture for resolving coreference in which deterministic resolution rules (or "sieves") are ordered from highest to lowest precision, and from lowest to highest recall. The advantages of this approach are similar to those of the previously-introduced rule-based architecture for entity and event extraction, including stability, interpretability by humans, and high overall performance.

However, though successful in the open domain, we discovered that the system proposed by Lee et al. (2013) is not well-suited to the biomedical domain, producing low-precision results due to overclustering. To account for this, we adapted the sieves to the biomedical domain by eliminating sieves that are redundant, uninformative in this domain, or insufficiently restrictive, as well as by creating new, domain-specific sieves that capitalize on domain knowledge.

For example, recognizing mutants (though the word *mutant* may not appear) will allow linking in sentences such as "... we prepared recombinant *H2AX-K134A*... The intensity of the band corresponding to histone H2AX methylation was significantly diminished in **the K134A mutant** compared with that of wild-type H2AX (H2AX-WT). . . .". Similarly, recognizing specific protein reactions allows otherwise difficult resolution, as in linking two dissimilar mentions of a single binding reaction in "*LL-37 forms a complex together with the IGF-1R* ... and **this binding** results in IGF-1R activation . . . ." We described this approach in detail in [12].

[12] Bell, D., G. Hahn-Powell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). An investigation of coreference phenomena in the biomedical domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC

2.6 Intrinsic evaluation: machine reading performance

Comparison with other reading systems

In an independently administered evaluation[13], Reach was found to extract signaling pathways at relatively high precision, at a throughput capable of reading the entire open source biomedical literature in short order. Participating systems extracted mechanistic information from 1,000 papers about the Ras signaling pathway over the course of a week. Two metrics were used to evaluate the participating systems:

[13] Conducted by MITRE in the DARPA Big Mechanism program (`www.darpa.mil/program/big-mechanism`). The publication cited above, which describes this evaluation, is in preparation. Please contact Matt Peterson (mpeterson@mitre.org) for early access to this document.

Precision This metric was calculated as the proportion of interactions that were considered "largely correct" in that (a) the interaction had to match the text evidence, (b) both participants (if present in the interaction) as well as the interaction type had to be correct, and (c) the negative information indicator (was the interaction negated or not in text?) had to be correct.

Throughput The estimated number of "largely correct" interactions produced from the 1,000 publications per day. Note that the correctness of entity grounding (i.e., linking the textual mentions of interaction participants to ids of actual entities in knowledge bases) was not a factor in calculating this precision measure. In this evaluation, throughput was used as a proxy for recall because true recall would be expensive to compute on such a large dataset.

Four other teams participated in the evaluation. The participating teams followed different approaches. For anonymity, we do not identify the participating consortia by name, but briefly describe their approaches. Team 1 implemented a pipeline of machine learning components that addressed various aspects of the task, such as identifying interaction types and interaction participants. Teams 2 and 3 implemented a hybrid approach, where they used machine learning to construct semantic representations of the text (Allen et al., 2008; Berant et al., 2013), and a rule-based component to extract domain-specific information from this open-domain semantic representation. Team 4 used a rule-based approach, with rules that focused solely on surface patterns. In this evaluation, Reach and Team 4 were part of the same consortium and evaluated jointly. The results are summarized in Table 2.6.[14]

Figure 2.8: Machine reading results in the Big Mechanism evaluation.

Team	*Throughput*	*Precision* (%)
Team 1	62	59
Team 2	342	23
Team 3	110	**63**
Reach + Team 4	**695**	49
Reach	486	62

The table shows that the Reach + Team 4 consortium obtains the best balance of precision and throughput, with the highest throughput and relatively high precision. Team 2 had the next highest throughput, but both its throughput and precision were more than twice as low as Reach + Team 4's scores. Teams 1 and 3 had higher precision scores, but their throughputs were considerably smaller: 11.2 and 6.3 times smaller than Reach + Team 4's throughput, respec-

[14] The precision scores in Table 2.6 are based on a slightly different sets of papers for each team. The motivation of this setup is that the number of interactions generated varied greatly across teams; the evaluation team had to score interactions from more papers to get reasonable precision numbers for submissions with fewer extractions. In particular, while all participants were scored on outputs from the same eight papers, Team 1, Team 3, and Reach + Team 4 were evaluated on two additional publications; Team 3 was further evaluated on three more.

tively.

While this evaluation reports results for Reach and Team 4 jointly, we aimed to tease out Reach's contribution in this consortium. To this end, we performed a post-hoc internal analysis of the data generated for this submission, separating the extractions produced by Reach from the extractions produced by Team 4. This analysis showed that Reach alone has a precision of 62% and is responsible for approximately 70% of the consortium throughput. These results, shown in the last row of Table 2.6, support the same observations: Reach has a throughput considerably higher than all the other teams, at precision approaching the highest precision value observed in the evaluation.

The high throughput observed for Reach has two explanations. First, the approach implemented in Reach, which includes Odin grammars that cover both syntax and surface patterns, coreference resolution, and other linguistic phenomena such as hedging (see Section), guarantees good coverage of the various linguistic phenomena encountered in this data. Second, the Reach grammar runtime system is fast: on average Reach processes a publication in less than five seconds. This allowed the team to easily process the entire dataset of 1,000 papers in the time allotted for this evaluation. In fact, the Reach submission was completed in the first few hours of the first of the seven days reserved for the evaluation.

All in all, this analysis demonstrates that Reach manages to maintain comparatively high precision without considerably sacrificing throughput. As we show in Section 2.7, this high throughput can be leveraged to increase precision by taking advantage of redundancy (i.e., the more times an interaction is extracted, the more likely it is to be correct).

Other biomedical tasks

Note that, while other efforts on extracting biomedical structures from free text certainly exist (Kim et al., 2009, 2011; Nédellec et al., 2013), they are not directly comparable to this work, for several reasons:

1. There are marked differences in task definitions behind Big Mechanism and other existing efforts. For example, the BioNLP 2009 dataset [15] contains six simple events, whereas Reach has 12. On the other hand, the BioNLP simple events include gene expression and transcription interactions, whereas Reach focuses strictly on post-translational modification (PTM) events. Further, the BioNLP datasets contain molecular-level regulation events, whereas Reach additionally extracts activation events that describe interactions at a higher abstraction level.

[15] 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 BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 1–9. Association for Computational Linguistics

2. There are considerable tokenization differences between BioNLP annotations and Reach. Specifically, BioNLP extracts subword events, e.g., where both the predicate and the corresponding argument are included in the same token, and subword arguments, e.g., where only part of a word is the argument of a predicate. For example, "phospho-p38" is labeled as an event in which the p38 protein is phosphorylated. Reach generally does not extract such subword events.

3. There are differences in how interactions were defined in BioNLP vs. Big Mechanism. For example, Binding (i.e., complex assembly) events in BioNLP can have an arbitrary number of arguments, whereas in Big Mechanism Binding events are binary (n-ary complex assembly interactions are represented as a sequence of binary Binding events).

4. Most BioNLP datasets contain only text from publication abstracts (Kim et al., 2009), or a mixture of abstracts and full publications, heavily biased towards abstracts (Kim et al., 2011). The BioNLP 2013 dataset (Nédellec et al., 2013) is the only one that contains solely text from full publications, but it is small (only 10 publications), which introduces a bias risk. In contrast, Reach was designed to robustly parse the full content of any biomedical paper.

Nevertheless, in order to put this work in a larger context, we implemented a simple comparative analysis in which we evaluated Reach on Phosphorylation events in the BioNLP 2013 dataset. Phosphorylation interactions are the most frequent PTM simple event in the BioNLP 2013 dataset, and they generally align well with the Reach definition. On the BioNLP 2013 development partition, Reach obtains a precision of 92.9%, a recall of 56.0%, and an F1 score of 69.9%.[16] In contrast, the Turku Event Extraction System (TEES) [17], the second best system in the BioNLP 2013 evaluation, obtained 83.9% precision and 83.5% recall on the development dataset. We find Reach's high precision encouraging, especially considering that Reach was never exposed to this dataset before this exercise, whereas all the other BioNLP participants used supervised learning, and tuned hyper parameters to maximize performance on this development partition.

[16] Using the approximate span and recursive criteria scorer, the standard scorer configuration in the BioNLP challenge.

[17] Björne, J. and T. Salakoski (2013). TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 16–25. Association for Computational Linguistics

To understand the lower recall, we inspected the false negatives (FN), i.e., phosphorylation events missed by Reach on the development partition of the BioNLP 2013 dataset. Our analysis confirms that these were caused by differences in task definition. In particular, 58% of the FNs were caused by subword events such as the one shown above in this subsection. Assuming Reach were modified to handle such subword events, its ceiling performance on phosphorylation events would be 92.9% precision, and 81.9% recall, for an F1 of 87.1% on the development partition, demonstrating the possibility of a score considerably higher than the one reported by the TEES system (83.9% precision and 83.5% recall). The other FNs were caused by faulty syntactic parsing (10%), misidentifying causes (7%), unhandled errors in the input such as "phosphor**ly**ation" (6%), latent arguments that are only supplied by domain knowledge (3%), and missing rule coverage (16%).

Lastly, please note that this difference in task definitions works both ways: Team 1 in Table 2.6 trained and tuned their components using the BioNLP datasets. As the table shows, this yielded low throughput compared to Reach in the Big Mechanism evaluation.

2.7 *Extrinsic evaluation: discovery of biological hypotheses*

This evaluation demonstrates that Reach-extracted pathway fragments improve the inference capacity of the Mutex algorithm, even when it already benefits from large curated models ("big mechanisms"). Specifically, we extended the Pathway Commons[18] human-curated pathways, which were used by Mutex in earlier work (Babur et al., 2015), with fragments extracted by Reach from all papers in the open access subset of PubMed[19]. Figure 2.7 compares the overlap of Pathway Commons with the information automatically extracted by Reach from this full-text subset of PubMed.

[18] `http://www.pathwaycommons.org/`

[19] 1,046,662 papers as of June 2015.

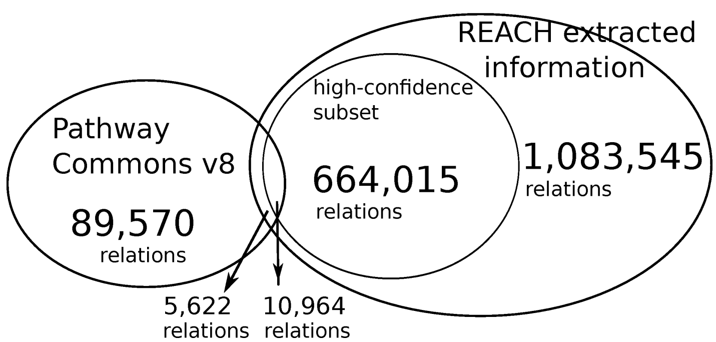


Figure 2.9: The Reach output is about 12 times larger than the size of Pathway Commons. We conjecture that the small overlap is caused by the fact that the Reach interactions are extracted from open-access publications, whereas Pathway Commons pathways come mostly from other, paywalled publications. The high-confidence subset is of relations that were found in more than one paper.

Using this combined prior network we were able to identify *previously unidentified*, but highly statistically significant mutually exclusively altered signaling modules in TCGA cancer datasets using the Mutex algorithm described above. Figure 2.7 and Figure 2.7 show Mutex groups for TCGA breast cancer, and Table 2.7 summarizes the findings for all enhanced cancer studies in TCGA. R represents the Mutex configuration using the combined Reach + Pathway Commons network, P the Mutex configuration using only the Pathway

Commons network, and W the Mutex configuration uninformed by any supporting network. In Table 2.7 we also include ablation results, e.g., **R**−P−W is the output of the R configuration without hypotheses discovered by either the P or W approaches. All in all, Table 2.7 highlights that machine reading is responsible for the discovery of new hypotheses in 7 cancers.

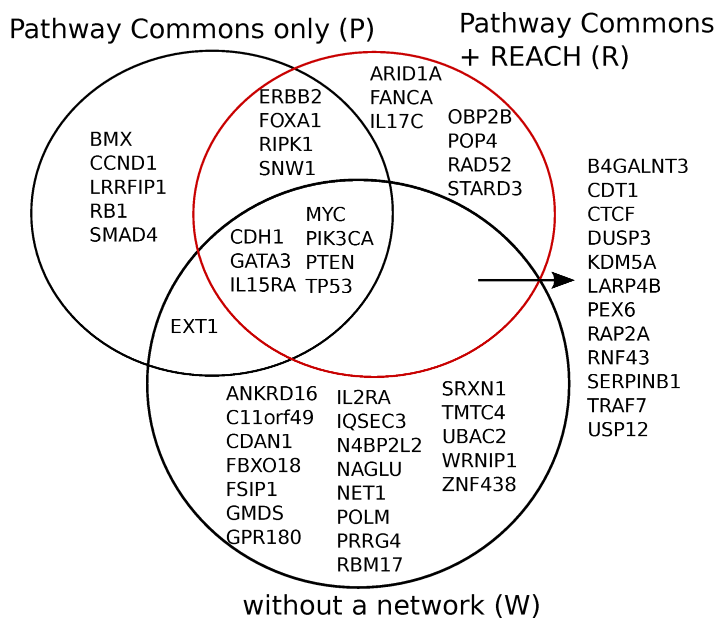


Figure 2.10: Reach allows Mutex to detect 7 new candidate "driver" genes for breast cancer which are not detected otherwise, when using Pathway Commons alone, or without using any network. We observed similar results for 6 other cancers in the TCGA dataset.

Error type	*Frequency*	*Example*	*Incorrect output*
Complex syntax	39% (7)	"In mouse models of leukemia and melanoma, IDH mutants accelerated cell cycle transition by activation of the MAPK and ERK pathway and repression of tumor suppressors CDKN2A and CDKN2B (Chaturvedi et al., Shibata et al.)"	CDKN2A controls IDH. The correct interaction to be extracted from this statement is: IDH controls CDKN2A.
NER	22% (4)	"At PND100, BPA significantly increased expression of EGFR ($p = 0.0132$), phospho-IGF-1R ($p = 0.007$), ..."	BPA is-a Protein. In this paper, BPA refers to the corresponding chemical not the protein with the same name.
Hedging	22% (4)	"Therefore, we next investigated whether CIC promotes mutant p53 GOF."	CIC activates p53, which is unsupported by the hedged statement.
Other	17% (3)	—	

Table 2.4: Correctness of the hypotheses generated by Mutex + Reach. The "With direction" column considers strict, directional hypotheses, e.g., *GATA3 activates PTEN*. The "Ignoring direction" column considers non-directional hypotheses, e.g., either *GATA3 activates PTEN* or *PTEN activates GATA3*.

A manual evaluation of these modules by an external cancer researcher (Table 2.7) reveals that, despite the inherent noise in machine reading, 65% of the hypotheses proposed by the Mutex algorithm

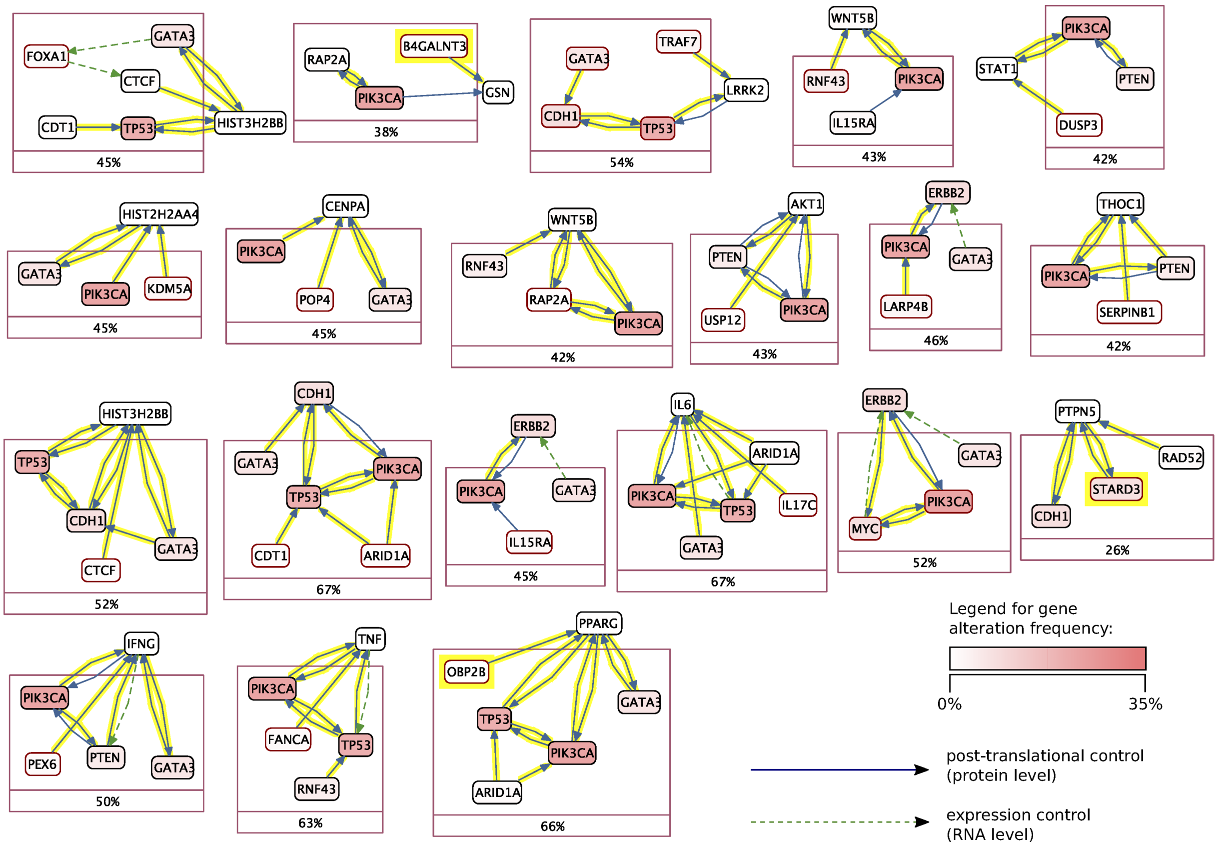


Figure 2.11: Mutex groups for TCGA breast cancer. This graph shows the interactions of the genes in each Mutex group and their targets. The highlighted relations exist in Reach data but not in Pathway Commons. Highlighted genes are not detectable without using Reach data.

Cancer study	**R**	P	W	**R**–P–W	**R**W–P
BLCA	2	2	6	0	0
BRCA	30	17	40	7	12
CESC	5	6	7	0	0
DLBC	0	5	0	0	0
GBM	23	14	40	3	7
HNSC	26	23	25	3	2
KICH	0	0	6	0	0
LAML	2	2	2	0	0
LGG	26	12	51	0	14
LIHC	12	17	16	0	0
LUAD	14	16	11	1	0
OV	7	11	7	2	0
PAAD	22	7	17	10	5
SARC	15	22	25	0	0
THCA	9	11	12	0	0
UVM	2	3	34	0	0

Figure 2.12: Mutex + Reach analysis of TCGA. The **R**–P–W and **R**W–P ablation experiments indicate that Reach extractions are responsible for the discovery of new hypotheses in 7 cancers.

	Hypotheses generated	*With direction*	*Ignoring direction*
Seen at least once	51	65%	71%
Seen at least twice	21	80%	80%

that had access to signaling pathways extracted by Reach are indeed correct according to the literature. Further, a simple redundancy filter that keeps Reach extractions only if they are seen at least twice in the literature increased this accuracy to 80%. This demonstrates that our approach systematically and incrementally increases coverage of prior, curated networks using NLP strategies, and, we believe, is valuable for molecular tumor boards and other cases where one needs to combine system-scale data with the knowledge in the literature.

However, a post-hoc error analysis of the incorrect hypotheses proposed by this approach (Table 2.7) indicates that machine reading is not a solved problem: 39% of the error are generated by incorrect syntactic analyses, 22% by incorrect entity recognition or grounding (e.g., in the example in the table "BPA" refers to the chemical Bysphenol A not the protein with the same name), and 22% are caused by hedged statements that were not supported by experimental results.

2.8 Conclusions and limitations

This work showed that large-scale machine reading of cancer literature marks a novel kind of cancer research that combines "big data" automatically extracted from the literature with "big mechanisms", i.e., large protein signaling pathways curated by domain experts.

We introduced Reach, a machine reading system that processes statements in the biomedical literature into mechanistic information. An independently-administered evaluation demonstrated that the proposed system outperforms other systems under a metric that combines precision and throughput. All in all, Reach achieved a relatively high precision at high throughput, capable of processing millions of articles within days. [20]

[20] The system is available as open-source software at `github.com/clulab/reach`.

We used Reach to process a large number of PubMed Central articles containing mechanistic information, and demonstrated that this information improves biological data analysis algorithms. Using a combination of information produced by Reach and Pathway Commons, we discovered new cancer driving mechanisms for seven cancers in the TCGA dataset. An external biologist who analyzed the hypotheses proposed by the algorithm found out that 65% of these are correct (i.e., they are supported by the literature). If we consider only interactions seen at least twice in the literature, 80% of the resulting hypotheses are correct.

Beyond the use case discussed here, this approach proposes a pipeline for information analysis in the biomedical domain that we believe generalizes beyond the domain addressed here. In this pipeline, machine reading is used to process a very large number of publications. This has the advantage of scalability beyond human capacity, but at the cost of introducing some noise. To mitigate the effect of such noise, biological data analysis algorithms (Mutex in this work) filter out the noise by identifying strong associations between machine data and patient data, and synthesize the information produced through machine reading into a small number of strong hypotheses.

Despite its strengths, the rules used here were manually crafted and the statements extracted were left unassembled. The next chapter will examine how these statements can be causally ordered to form longer pathways with greater explanatory power, while Chapter 5 will show how similar rules can be learned automatically.

3 Causal precedence of events

3.1 Overview

This chapter describes an approach first introduced in Hahn-Powell et al. (2016)[1] for the ordering of causal statements automatically extracted from the literature. Causal precedence between biochemical interactions is crucial in the biomedical domain, because it transforms collections of individual interactions, e.g., bindings and phosphorylations, into the causal mechanisms needed to inform meaningful search and inference. Here, we analyze *causal* precedence in the biomedical domain as distinct from open-domain, *temporal* precedence. First, we describe a novel, hand-annotated text corpus of causal precedence in the biomedical domain. Second, we use this corpus to investigate a battery of models of precedence, covering rule-based, feature-based, and latent representation models. The highest-performing individual model achieved a micro F1 of 79 points, approaching the best performers on the simpler temporal-only precedence tasks. Feature-based and latent representation models each outperform the rule-based models, but their performance is complementary to one another. We apply a sieve-based architecture to capitalize on this lack of overlap, achieving a micro F1 score of 77 and recall of 90%. The next chapter explores the meaning of assembly in a diverse environment comprised of multiple domains, and presents of means of

[1] Hahn-Powell, G., D. Bell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). This before That: Causal Precedence in the Biomedical Domain. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pp. 146–155. Association for Computational Linguistics. DOI: 10.18653/v1/W16-2920

3.2 Introduction

In the biomedical domain, an enormous amount of information about protein, gene, and drug interactions appears in the form of natural language across millions of academic papers. There is a tremendous ongoing effort (Nédellec et al., 2013; Kim et al., 2012, 2009) to extract individual chemical interactions from these texts, but these interactions are only isolated fragments of larger causal mechanisms such as protein signaling pathways. Nowhere, however, including any database, is the complete mechanism described in a form that lends itself to causal search or inference. The absence of such a database is not for lack of trying; Pathway Commons [2] aims to address the need, but its authors estimate it currently covers 1% of the literature due to the high cost of annotation[3]. This issue only grows more pressing with the yearly growth in biomedical publishing, which presents an otherwise insurmountable challenge for biomedical researchers to query and interpret.

[2] Cerami, E. G., B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, pp. D685–D690. DOI: 10.1093/nar/gkq1039

[3] Personal communication.

The Big Mechanism program (Cohen, 2015) aims to construct exactly such large-scale mechanistic information by reading and assembling protein signaling pathways that are relevant for cancer, and exploit them to generate novel explanatory and treatment hypotheses. Although prior work has addressed the challenging area of temporal (Chambers et al., 2014) precedence in the open domain, the biomedical domain presents very different data and, consequently, requires novel techniques. Precedence in mechanistic biology is *causal* rather than *temporal*. Though event temporality is crucial to understanding electronic health records for individual patients (Bethard et al., 2015, 2016), its contribution to the understanding of biomolecular reactions is less clear as these events and processes may repeat in extremely short cycles, continue without end, or overlap in time. At any level of abstraction, causal precedence encodes mechanistic information and facilitates inference over spotty evidence. For the purpose of this work, *precedence* is defined for two events, A and B, as

[4] Mihăilă, C., T. Ohta, S. Pyysalo, and S. Ananiadou (2013). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, **14**(1), pp. 1–18. ISSN 1471-2105. DOI: 10.1186/1471-2105-14-2

Very little annotated data exists for causal precedence, especially efforts focusing on signaling pathways. BioCause [4], for instance, is centered on connections between claims and evidence and contains only 51 annotated examples of causal precedence[5]. This work[6] offers three contributions in aid of automatically extracting causal ordering in biomedical text. First, we provide and describe a dataset of real text examples, manually annotated for causal precedence. Second, we analyze the efficacy of a battery of different models in automatically determining precedence, built on top of the Reach automatic reading system (Valenzuela-Escárcega et al., 2015a,b; Valenzuela-Escárcega et al., 2017) and measured against this novel corpus. In particular, we investigate three classes of models: (a) deterministic rule-based models inspired by the precedence sieves proposed by [7], (b) feature-based models, and (c) models that rely on latent representations such as long short-term memory (LSTM) networks [8]. Our analysis indicates that while independently the top-performing model achieves a micro F1 of 79, these models are largely complementary with a combined recall of 90 points. We follow the presentation of results with an error analysis of these models to motivate and inform future research.

[5] These are marked in the BioCause corpus as `Causality` events with `Cause` and `Effect` arguments. The remaining 800 annotations are claim-evidence relations.

[6] The corpus, tools, and system introduced in this work are publicly available at `https://github.com/myedibleenso/this-before-that`

[7] Chambers, N., T. Cassidy, B. McDowell, and S. Bethard (2014). Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association of Computational Linguistics*, **2**, pp. 273–284

[8] Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735

Relation	*Example*
E1 precedes E2	A is phosphorylated by B. Following its phosphorylation, A binds with C.
E2 precedes E1	A is phosphorylated by B. Prior to its phosphorylation, A binds with D.
Equivalent	The phosphorylation of A by B. A is phosphorylated by B.
E1 specifies E2	A is phosphorylated by B at Site 123. A is phosphorylated by B.
E2 specifies E1	A is phosphorylated by B. A is phosphorylated by B at Site 123.
Other	B does not regulate C when C is bound to A.
None	A phosphorylates B. A ubiquitinates C.

Table 3.1: The seven inter-event relation labels annotated in the corpus. The "precedes" labels are causal. Subsumption is captured with the "specifies" labels.

3.3 Related work

Prior work has largely focused on open-domain texts. Bethard and Martin (2008) constructed a corpus of temporal and causal relations and then trained classifiers for relation identification. Riaz and Girju (2013) classified verb pairs by degree of causal association. Mirza and Tonelli implemented a hybrid sieve-based approach for causal relation detection between events that includes a set of causal verb rules and corresponding syntactic dependencies and a feature-based classifier [9]. Though focused on temporal ordering, Chambers et al. (2014) adopt a sieve-based approach with high-precision, deterministic sieves preceding and constraining lower-precision, higher-recall machine learning sieves. As with our system, the deterministic sieves were linguistically motivated, and had the additional advantage of operating over time expressions (*during*, *Friday*, etc.) as well as events, the former of which are typically lacking in the biomedical domain. To our knowledge, we are the first to investigate causal precedence in the biomedical domain.

[9] Mirza, P. (2016). *Extracting Temporal and Causal Relations between Events*. Ph.D. thesis, University of Trento; Mirza, P. (2014). Extracting Temporal and Causal Relations between Events. In *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10–17. Association for Computational Linguistics. DOI: 10.3115/v1/P14-3002; and Mirza, P. and S. Tonelli (2016). CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 64–75. The COLING 2016 Organizing Committee

3.4 A corpus for causal precedence in the biomedical domain

Our corpus annotates several types of relations between mentions of biochemical interactions. Following common terminology promoted by the BioNLP shared tasks, we will interchangeably use "events" to refer to these interactions. To generate candidate events for our planned annotations, we ran the Reach event extraction system (Valenzuela-Escárcega et al., 2015a,b) over the full text[10] of 500 biomedical papers taken from the Open Access subset of PubMed[11]. The events extracted by Reach are biochemical events of two types: simple events such as phosphorylation that modify one or more entities (typically proteins), and nested events (regulations) that have other events as arguments.

[10] We chose to ignore the "references", "materials", and "methods" sections, which generally do not contain mechanistic information.

[11] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

To improve the likelihood of finding pairs of events with a relevant link, we filtered event pairs by imposing the following requirements for inclusion in the corpus:

1. *Event pairs must share at least one participant.* This constraint is based on the observation that interactions that share participants are more likely to be connected.

2. *Event pairs must be within 1 sentence of each other.* Similarly, discourse proximity increases the likelihood of two events being related.

3. *Event pairs must not share the same type.* This helps to maximize the diversity of the dataset.

4. *Event pairs must not already be contained in an extracted Regulation event.* For example, we did not annotate the relation between the binding and the phosphorylation events in "The binding of X and Y is inhibited by X phosphorylation", because it is already captured by most state-of-the-art biomedical event extraction systems.

After applying these constraints, only 1700 event pairs remained. In order to rapidly annotate the event pairs, we developed a browser-based annotation UI that is completely client-side (see Figure 3.4). Using this tool, we annotated 1000 event pairs for this work; 84 of these were discarded due to severe extraction errors. The annotations include the event spans, event triggers (i.e., the verbal or nominal predicates that indicate the type of interaction such as "binding" or "phosphorylated"), source document, minimal sentential span encompassing both event mentions, and whether or not the event pair involves coreference for either the event trigger or the event participants. For events requiring coreference resolution, we expanded the encompassing span of text to also capture the antecedent. Note that domain-specific coreference resolution is a component of the event extraction system used here [12].

[12] Bell, D., G. Hahn-Powell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). An investigation of coreference phenomena in the biomedical domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC

When describing the relations between these event pairs, we refer to the event that occurs first in text as Event 1 (E1) and the event that follows as Event 2 (E2). Each (E1, E2) pair was assigned one of seven

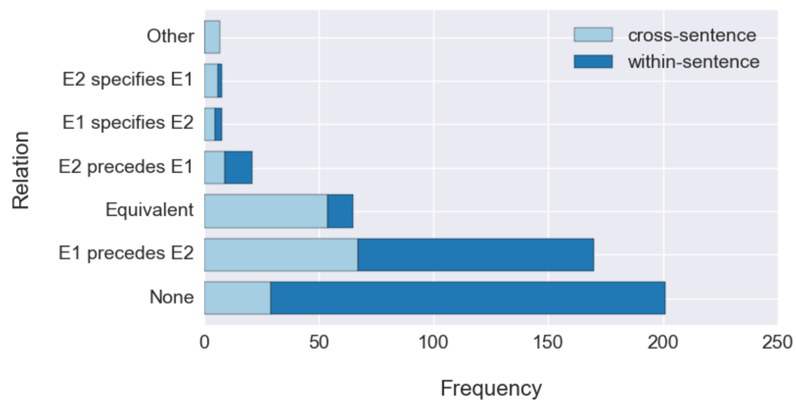


Figure 3.1: The distribution of assembly relation labels both within and across sentences.

labels: "E1 precedes E2", "E2 precedes E1", "Equivalent", "E1 specifies E2", "E2 specifies E1", "Other", or "None". Table 3.4 provides examples for each of these labels. We converged on these labels because they are fundamental to the assembly of causal mechanisms from a collection of events. Collectively, the seven labels address three important assembly tasks: *equivalence*, i.e., understanding that two event mentions discuss the same event, *subsumption*, i.e., the two mentions discuss the same event, but one is more specific than the other, and, most importantly, *causal precedence*, the identification of which is the focus of this work. During the annotation process, we came across examples of other relevant phenomena. We grouped these instances under the label "Other" and leave their analysis for future work.

Though simplified, the examples in Table 3.5 illustrate that this is a complex task sensitive to linguistic evidence. For example, the direction of the precedence relations in the first two rows in the table changes based on a single word in the context ("prior" vs. "following").

In terms of the distribution of relations, causal precedence pairs appear more frequently within the same sentence, while cases of the subsumption ("specifies") and equivalence relations are far more common across sentences (see Figure 3.4). Coreference is involved in 10–15% of the instances for each relation label (see Figure 3.4).

The annotation process was performed by two linguists familiar with the biomedical domain. To minimize errors, the annotation task was initially performed together at the same workstation.[13] On a randomly selected sample of 100 event pairs, the two annotators had a Cohen's kappa score [14] of 0.82, indicating "almost perfect" agreement for the *precedes* labels [15].

[13] Similar to pair programming.

[14] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), pp. 37–46. DOI: 10.1177/001316446002000104

[15] Landis, J. R. and G. G. Koch (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**(1), pp. 159–174. ISSN 0006341X, 15410420. DOI: 10.2307/2529310

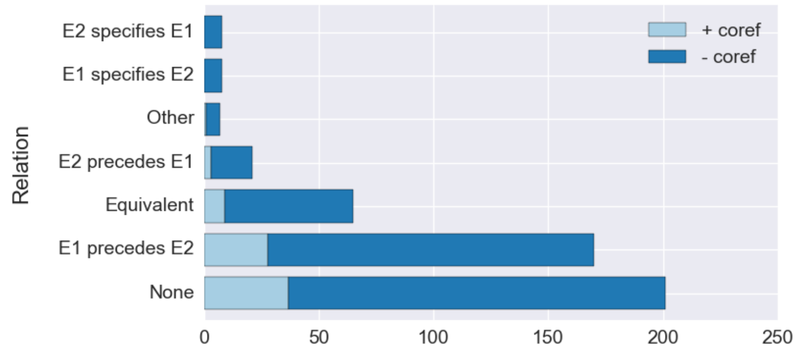


Figure 3.2: The distribution of event pairs involving coreference across assembly relations.

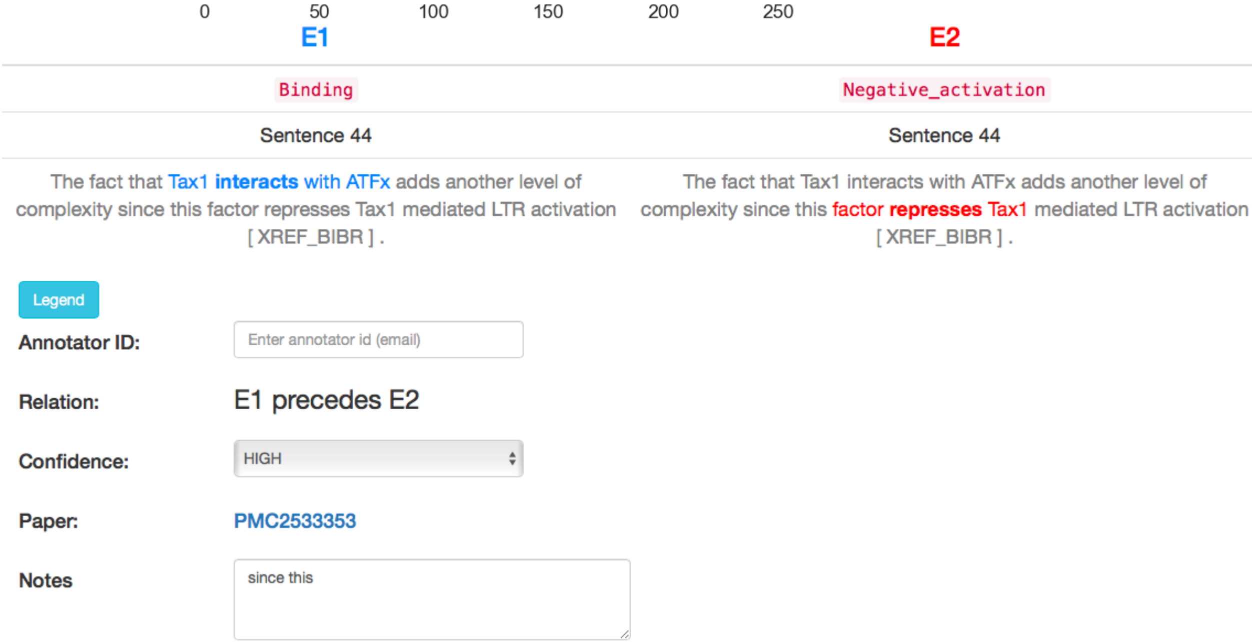


Figure 3.3: Browser-based tool for annotating assembly relations in text. An annotation instance consists of a pair of event mentions. The annotator assigns a label to each pair of events using the number keys and navigates from annotation to annotation using the arrow keys. E1 refers to the event in the pair that appears first in the text. The event span is formatted to stand out from the surrounding text. The "Paper" field provides the annotator with easy access to the full text of the source document for the current annotation instance. Annotations can be exported to JSON and reloaded via a local storage cache or through file upload.

3.5 *Models of causal precedence*

We have developed both deterministic, interpretable models and automatic, machine-learning models for detecting causal precedence in our dataset. Importantly, the models covered in this work focus solely on causal precedence, which is the most complex relation annotated in the dataset previously introduced. Thus, for all experiments discussed here, we reduce these annotations to three labels: "E1 precedes E2", "E2 precedes E1", and `Nil`, which covers all the other labels in the corpus.

Model	*Rules*
Intra-sentence	54
Inter-sentence	29
Reichenbach	8

Table 3.2: Few rules defined each deterministic model of precedence compared with the number of features for the machine learning models.

Deterministic models

The deterministic models are defined by a small number of hand-written rules using the Odin event extraction framework [16]. The number of rules for each model is shown in Table 3.5, and sharply contrast with the 92,711 features introduced later (Table 3.5) that are used by our machine-learning models. In order to avoid overfitting, all of the deterministic models were created without reference to the annotation corpus, using general linguistic expertise and domain knowledge.

[16] Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2015c). Description of the Odin Event Extraction Framework and Rule Language

Intra-sentence ordering Within sentences, syntactic regularities can be exploited to cover a large variety of grammatical constructions indicating precedence relations. Rules defined over dependency parses [17] capture precedence in sentences like those in (1) and (2) as well as many others.

[17] de Marneffe, M.-C. and C. D. Manning (2008). The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Coling 2008 Organizing Committee

(1) [The RBD of PI3KC2B binds HRAS]$_{\text{after}}$, when [HRAS is not bound to GTP]$_{\text{before}}$

(2) [The ubiquitination of A]$_{\text{before}}$ is followed by [the phosphorylation of B]$_{\text{after}}$

Other phrases captured include: "precedes", "due to", "leads to", "results in", etc.

Inter-sentence ordering Although syntax operates over single sentences, cross-sentence time expressions can indicate ordering, as shown in Examples (3) and (4). We exploit these regularities as well by checking for sentence-initial word combinations.

(3) [A is phosphorylated by B]$_{\text{before}}$. As a downstream effect, [C is . . .]$_{\text{after}}$

(4) [A is phosphorylated by B]$_{\text{before}}$. [C is then . . .]$_{\text{after}}$

Other phrases captured include: "Later", "In response", "For this", and "Ultimately".

Verbal tense- and aspect-based (Reichenbach) ordering Following [18], we use deterministic rules to establish precedence between events that have certain verbal tense and aspect. These rules are derived from

[18] Chambers, N., T. Cassidy, B. McDowell, and S. Bethard (2014). Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association of Computational Linguistics*, **2**, pp. 273–284

linguistic analysis of tense and aspect by [19]. Example (5) illustrates a case in which we can accurately infer order just from this information. Because *has been phosphorylated* has past tense and perfective aspect, this model concludes that it precedes *share* (present tense, simple aspect) and thus the binding of histone H2A.

(5) These [PTIP] proteins also share the ability to bind histone H2A (or H2AX in mammals) that has been phosphorylated. . . .

[19] Reichenbach, H. (1947). *Elements of Symbolic Logic*. Macmillan. ISBN 0486240045; and Derczynski, L. and R. Gaizauskas (2013). Empirical Validation of Reichenbach's Tense Framework. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pp. 71–82. Association for Computational Linguistics

The logic determining which tense-aspect combinations receive which precedence relations is identical to CAEVO, which is possible because it is open source[20]. However, CAEVO operates over annotations that include gold tense and aspect values, whereas this model additionally detects tense and aspect using Odin rules before applying this logic.

[20] `https://github.com/nchambers/caevo`

Feature-based models

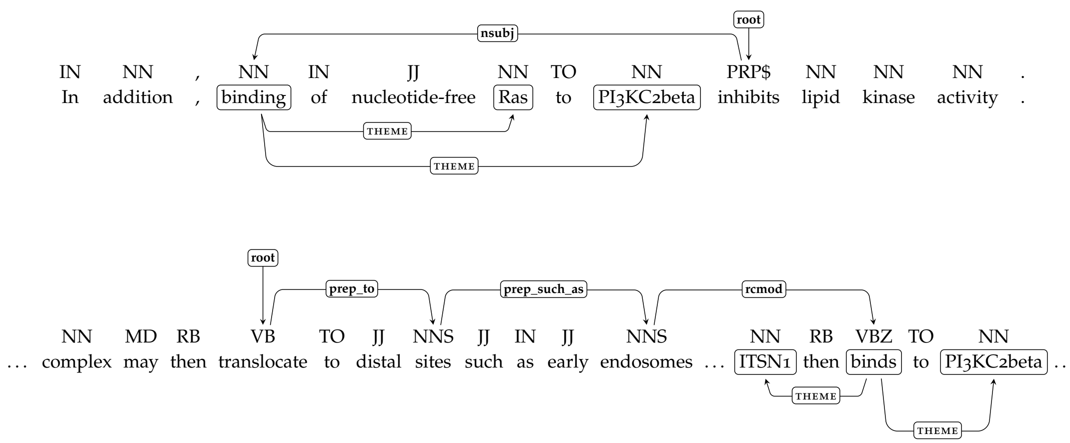
Most instances of causal precedence cannot be captured with deterministic rules, because they lack explicit words, phrases, or syntactic structures that unambiguously mark the relation. Using a combination of the surface, syntactic, and taxonomic features outlined in Table 3.5, we trained a set of statistical classifiers to detect causal precedence relations between pairs of events in our corpus. For training and testing purposes, we treated any instance not labeled as either "E1 precedes E2" or "E2 precedes E1" as a negative example. We examined the following statistical models: a linear kernel SVM (Chang and Lin, 2011), logistic regression (Fan et al., 2008), and random forest[21] (Surdeanu et al., 2014). For the SVM and logistic regression (LR) models, we also compared the effects of L1 and L2 regularization.

[21] Abbreviated as RF

	Feature	*Description*
Event	Event labels	The taxonomic labels Reach assigned to the event (e.g. *phosphorylation* $\rightarrow$ Phosphorylation, AdditiveEvent, ...).
	Event trigger	The predicate signaling an event mention (ex. "phosphorylated", "phosphorylation").
	Event trigger + label	A concatenation of the event's trigger with the event's label.
	token *n*-grams with entity replacement	*n*-grams of the tokens in the mention span, where each entity is replaced with the entity label (ex. "the ABC protein" $\rightarrow$ "the PROTEIN"). If an entity is shared between pairs of events, replace it with the label SHARED.
	token *n*-grams with role replacement	*n*-grams of the tokens in the mention span, where each argument is replaced with the argument role (ex. "A inhibits the phosphorylation of B" $\rightarrow$ "CONTROLLER inhibits the CONTROLLED")
	Syntactic path from trigger to args	Variations of the syntactic dependency path from an event's trigger to each of its arguments (unlexicalized path, path + lemmas, trigger $\rightarrow$ argument role, trigger $\rightarrow$ argument label, etc.).
Event-Event (surface)	Interceding tokens (*n*-grams)	*n*-grams (1-3) of the tokens between E1 and E2.
Event-Event (syntax)	Cross-sentence syntactic paths	A concatenation of the syntactic path from the sentential ROOT to an event's trigger (see the example in Figure 3.5).
	Trigger-to-trigger syntactic paths (within sentence)	the syntactic path from the trigger of E1 to the trigger of E2
	Shortest syntactic paths	The shortest syntactic path between E1 and E2 (restricted to intra-sentence cases).
	Syntactic distance	The length of each syntactic path (restricted to intra-sentence cases).
Coreference	*Event* features for anaphors	Whether or not an event mention is resolved through coreference. For cases of coreference, generate the *Event* features prefixed with "coref-anaphor" for the text labeled "E1-anaphor" in the following example: (6) [A binds with B]$_{\text{E1-antecedent}}$ (7) [This interaction]$_{\text{E1-anaphor}}$ precedes the [phosphorylation of C]$_{\text{E2}}$
	Resolved arguments	Which arguments, if any, were resolved through coreference. For example: [The mutant$_{\text{THEME}}$ binds with B$_{\text{THEME}}$]$_{\text{E1}}$ $\rightarrow$ `theme:resolved`

Table 3.3: An overview of the primary features used in the feature-based classifier, grouped into four classes: *Event* – features extracted from the two participating events, in isolation; *Event-Event (surface)* – features that model the lexical context between the two events; *Event-Event (syntax)* – features that model the syntactic context between the two events; and *Coreference* – features that capture coreference resolution information that impact the participating events.

In addition, **binding of nucleotide-free Ras to PI3KC2β** inhibits its lipid kinase activity. The PI3KC2β and Ras complex may then translocate to distal sites such as early endosomes (EE) where **ITSN1 then binds to PI3KC2β** leading to the release of nucleotide-free Ras and activation of the lipid kinase activity of PI3KC2β.



ROOT >NSUBJ + ROOT >PREP_TO >PREP_SUCH_AS >RCMOD

Figure 3.4: Generation procedure for the cross-sentence syntactic path feature. For each event in a pair, we find the shortest syntactic path originating from the sentential root node leading to a token in the event's trigger. The two syntactic paths are then joined using the + symbol to form a single feature.

Latent representation models

Due to the complexity of the task and variety of causal precedence instances encountered during the annotation process, it is unclear whether a linear combination of engineered features is sufficient for broad coverage classification. For this reason, we introduce a latent feature representation model using an LSTM (Hochreiter and Schmidhuber, 1997; Bergstra et al., 2010; Chollet, 2015) to capture underlying semantic features by incorporating long-distance contextual information and selectively persisting memory of previous event pairs to aid in classification.

The basic architecture is shown in Figure 3.5. The input to this model is the provenance of the relation, i.e., the whole text containing the two events and the text in between. Formally, this is represented as a concatenated sequence of 200 dimensional vectors where each vector in the sequence corresponds to a token in the minimal sentential span encompassing the event pair being classified. Intuitively, this LSTM "reads" the text from left to right and outputs a classification label from the set of three when done. We consider two variations of this model: the basic model (LSTM) with the vector weights for each token uninitialized and a second form (LSTM+P) where the vectors are initialized using pre-training. In the pre-training configuration, the vector weights are initialized using word embeddings generated by a word2vec (Mikolov et al., 2013a; Řehůřek and Sojka, 2010) model trained on the full text of over 1 million biomedical papers taken from the Open Access subset of PubMed. Because the corpus is only 1000 annotations, it was thought that pre-training could improve prediction of causal precedence and guide the model with distributional semantic representations specific to this domain.

Building on this simple blueprint, we designed a three-pronged "pitchfork" (FLSTM) where the span of E1, the span of E2, and the minimal sentential span encompassing E1 and E2 each serve as a separate input, allowing the model to explicitly address each of them as well as discover how these three inputs relate to one another. This architecture is shown in Figure 3.5. Each input feeds into its own LSTM and corresponding dropout layer before the three forks are merged via a concatenation of tensors. Like the basic model, one version of the "pitchfork" is trained with vector weights initialized using the pre-trained word embeddings (FLSTM+P).

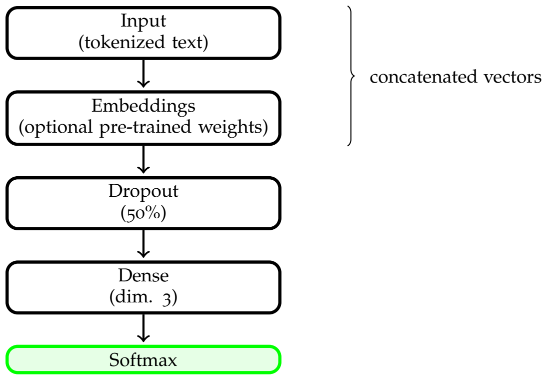


Figure 3.5: Architecture for the basic latent feature model using the minimal sentential span encompassing events 1 and 2 as input.

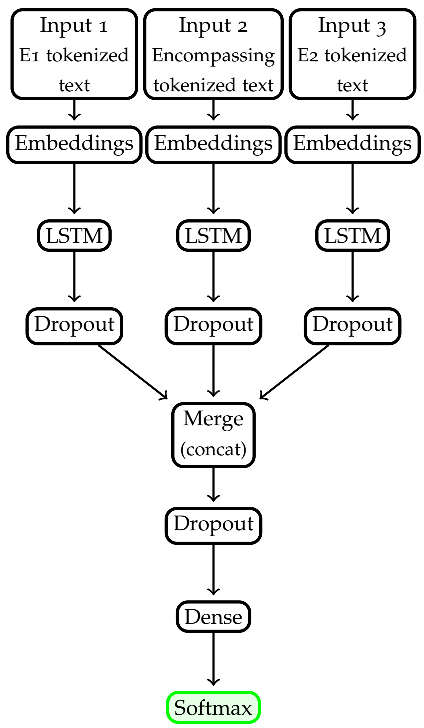


Figure 3.6: Modified architecture for a latent feature model with three-pronged input: the text of event 1 (left), the minimal sentential span encompassing events 1 and 2 (middle), and the text of event 2 (right).

3.6 Results

We summarize the performance of all these models on the dataset previously introduced in Table 3.6. We report results using micro precision, recall, and F1 scores for each model. With fewer than 200 instances of causal precedence occurring in 1000 annotations, training and testing for both the feature-based classifiers and latent feature models was performed using stratified 10-fold cross validation. For the latent feature models, training was parameterized using a maximum of 100 epochs with support for early stopping through monitoring of validation loss[22]. Weight updates were made on batches of 32

[22] The validation set used for each fold came from a different class-balanced fold.

examples and all folds completed in fewer than 50 epochs.

The table also includes a sieve-based ensemble system, which performs significantly better than the best-performing single model. In this architecture, the sieves are applied in descending order of precision, so that the positive predictions of the higher precision sieves will always be preferred to contradictory predictions made by subsequent, lower-precision sieves. As sieves are added, the F1 score remains fairly constant, while recall increases at the cost of precision.

	Model	*P* (%)	*R* (%)	*F1* (%)
Rules	Causal patterns	**85**	56	68
	Reichenbach	0	0	0
Feature-based	LR+L1	79	76	78
	LR+L2	79	75	77
	SVM+L1	80	78	**79**
	SVM+L2	77	74	75
	RF	77	80	78
Latent	LSTM	75	73	74
	LSTM+P	77	75	76
	FLSTM	78	76	77
	FLSTM+P	**83**	74	78
Sieves	Combined (all)	67	**90**	77

Table 3.4: Results of all proposed causal models, using stratified 10-fold cross-validation. The combined system is a sieve-based architecture that applies the models in decreasing order of their precision. The combined system significantly outperforms the best single model, SVM with L1 regularization, according to a bootstrap resampling test ($p = 0.022$).

Despite some obvious patterns noted in Table 3.4, the deterministic models perform the worst due in large part to their rarity in the corpus. An analysis of this result is given in Section 3.7. Overall, our top-performing model was the linear kernel SVM with L1 regularization. In all cases, the feature-based classifiers outperform the latent feature representations, suggesting that in cases such as this where little data is available, feature-based classifiers capitalizing on high-level linguistic features are able to better generalize and thus outperform latent feature models. However, as our discussion in Section 3.7 will show, our combined model demonstrates that the latent and feature-based models are largely complementary.

3.7 *Discussion*

Overall, results are promising, particularly in light of the conscious choice to omit (causal) regulation reactions from this task, as they are already captured by the Reach reading system.

However, the deterministic models created so far have extremely low recall, such that it is difficult even to determine their precision. An analysis of the Reichenbach model reveals one source of this low coverage. In short, although writers *could* describe causal mechanisms using temporal indicators such as tense and aspect, temporal description is rare enough in this domain not to be represented in our randomly sampled database. Table 3.7 illustrates the lack of overlap with informative tense-aspect combinations; a single tense is used per passage, and no perfective aspect is used.

E1↓, E2→		*past*		*pres.*		*fut.*	
		simple	*perf.*	*simple*	*perf.*	*simple*	*perf.*
past	*simple*	69	0	38	0	0	0
	perf.	0	0	0	0	0	0
pres.	*simple*	49	0	134	0	1	0
	perf.	0	0	0	0	0	0
fut.	*simple*	0	0	0	0	0	0
	perf.	0	0	0	0	0	0

Table 3.5: Tense and aspect for events containing verbs in the present study. Highlighted cells are tense-aspect combinations that are informative for establishing temporal precedence, following Chambers et al. (2014). All but one event pair fall outside of these informative combinations, and that exceptional pair was a false positive case.

Similarly, the time expressions required by the deterministic intra- and inter-sentence precedence rules are rare enough to make them ineffective on this sample.

Model overlap

As Chambers et al. (2014), Mirza (2016), and many other algorithms have shown, models can be applied sequentially in "sieves" to produce higher-quality output. Ideally, each model in a sieve-based system will capture different portions of the data through a mixture of approaches, distinguishing this method from more naive ensembles in which the contributions of a lone component would be washed out.

Error analysis

We performed an analysis of the false positives shared by all feature-based classifiers, in addition to the false negatives shared by all models. Here we limit our discussion to only the most prominent characteristic shared by the majority of false positives.

Discourse information More than half of the false positives share contrastive discourse features, suggesting that a model of discourse could improve classifier discrimination. Example (8) demonstrates such a contrastive structure, which *whereas* introduces a clause (and event) that is contrasted and therefore both temporally and causally distinct from the following clause (and event). The existence of regular cues like *whereas* indicates that a feature to explicitly model these structures is possible.

(8) Whereas [PRAS40 inhibits the mTORC1 activity via raptor]$_{E1}$, DEPTOR was identified to interact directly with mTOR in both [mTORC1 and mTORC2 complexes]$_{E2}$

3.8 Conclusions and limitations

These are the first experiments regarding automatic annotation of causal precedence in the biomedical domain. Although the dearth of temporal expressions and other regular linguistic cues make the task especially difficult in this domain, the initial results are promising, and demonstrate that a sieve-based system of the models tested here improves performance over the top-performing individual component. Both the annotation corpus and the models described here represent large steps toward linking automatic reading to a larger, more informative biological mechanism.

4 A platform for multi-domain scientific discovery

"Knowledge can be public, yet undiscovered, if independently created fragments are logically related but never retrieved, brought together, and interpreted."

Don R. Swanson (1986b)

4.1 Overview

This chapter introduces a modular approach for literature-based discovery first introduced in Hahn-Powell et al. (2017) [1], which consists of a machine reading and knowledge assembly component that together produce a graph of influence relations (e.g., "A promotes B") from a collection of publications. A search engine is used to explore direct and indirect influence chains. Query results are substantiated with textual evidence, ranked according to their relevance, and presented in both a table-based view, as well as a network graph visualization. This platform for scientific discovery facilitates the rapid review of literature, as well as the collaborative construction of conceptual models supported by findings mined from text. The approach operates in both domain-specific settings, where there are knowledge bases and ontologies available to guide reading, and in multi-domain settings where such resources are absent. We demonstrate that this deep reading and search system reduces the effort needed to uncover "undiscovered public knowledge", and that with the aid of this tool a domain expert was able to drastically reduce her model building time from months to two days.

[1] Hahn-Powell, G., M. A. Valenzuela-Escárcega, and M. Surdeanu (2017). Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Software Demonstrations*, pp. 103–108. ACL. doi: 10.18653/v1/P17-4018

4.2 Introduction

The extreme growth in the rate of scientific publications since the 1900s has created the need for large-scale text mining and machine reading approaches capable of understanding the claims and assertions made by authors. Although a number of machine reading systems have been developed to mine individual facts (e.g., biochemical interactions) from these publications, there is limited work on assembling and interpreting these fragments. This limitation may result in solutions to critical problems being overlooked, as many tasks today cross several disciplines that interact only minimally. Swanson [2] described this problem as "undiscovered public knowledge".

This chapter describes a machine reading and assembly approach that facilitates connections between different research efforts and research communities across several fields. Following past efforts in literature-based discovery [3], we introduce a modular system that performs (a) information extraction and assembly from publications, and (b) hypothesis exploration using a custom search engine that queries the knowledge graph of direct and indirect links uncovered by the machine.

We elected to focus on links that highlight influence relations between two concepts (e.g., "CTCF activates FOXA1") in the biomedical domain. Importantly, our approach reads statements about influence relations from publications, and does not attempt to verify these findings directly through separate modeling.[4] We apply our system to build influence graphs for two scenarios: (a) biomolecular explanations of cell signaling pathways with applications to cancer research (this is a single domain rich with knowledge base (KB) resources to guide information extraction), and (b) factors influencing children's health (this task crosses multiple domains, and has limited supported from KBs).

[2] Swanson, D. R. (1986b). Undiscovered public knowledge. *The Library Quarterly: Information, Community, Policy*, **56**(2), pp. 103–118

[3] Smalheiser, N. R. and D. R. Swanson (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, **57**(3), pp. 149 – 153. ISSN 0169-2607. DOI: https://doi.org/10.1016/S0169-2607(98)00033-9; and Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, **3**(1), p. 2. ISSN 1742-5581. DOI: 10.1186/1742-5581-3-2

[4] In other words, this work assumes that the authors' statements are correct. Although these statements often contain causal language (e.g., "A causes B"), care is taken to avoid referring to these relations as causal, since the approach outlined here does not attempt to verify quantitative findings directly.

4.3 *Related work*

Machine reading is an extension of information extraction (IE), which encompasses a series of techniques for mining unstructured data (e.g., text) and "extracting limited kinds of semantic content" (Jurafsky and Martin, 2018, Chapter 21) with the aim of completing or enriching some knowledge base (KB) centered around entities of interest. The nature and scope of this semantic content and the definition of entity is determined by the application. Tasks can be diverse as identifying proteins and their reported interactions (Kim et al., 2009; Hirschman et al., 2005; Kim et al., 2011; Ohta et al., 2013), interpreting radiology reports (Mykowiecka et al., 2009), understanding political events as they are reported in the media (Afolabi et al., 2008), or analyzing patents to functions and properties which define technological trends (Yoo, 2012). A commonly used example is the task of completing Wikipedia infoboxes (Wu et al., 2008) by analyzing the text of the accompanying article (see Figure 4.3 for an example).

Approaches to information extraction and text data mining can be broadly classified into three approaches: statistical, rule-based, and

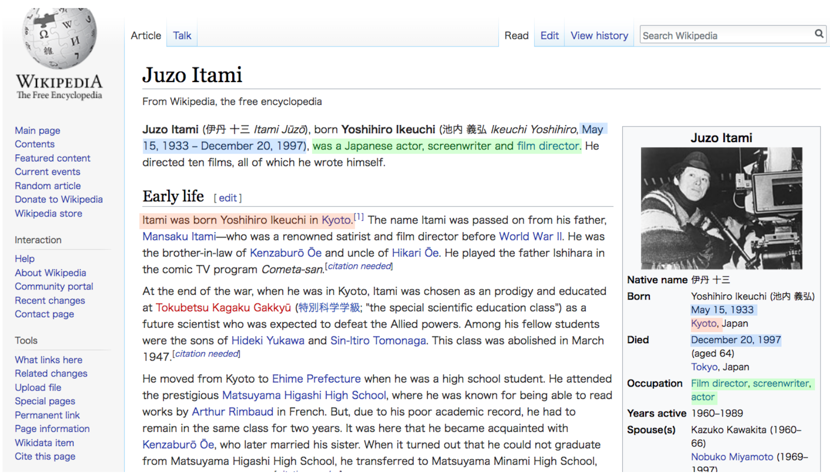


Figure 4.1: An example of Infobox completion for the Wikipedia page on Itami Juzo. Spans of text from the accompanying article which could be used to populate the infobox are highlighted in a color that corresponds to its target field.

hybrid models.

Statistical approaches Dominant approaches to information extraction use statistical models where the representation of the input text are hand-crafted features, latent representations, or some combination (Toutanova and Chen, 2015). Many models leverage knowledge bases to guide extraction (Toutanova et al., 2015).

Rule-based approaches The earliest approaches to information extraction were rule-based systems (Hearst, 1992; Appelt et al., 1993; Riloff, 1996; Appelt and Onyshkevych, 1998). [5] The generalizabilty and expressivity of such patterned-based approaches is largely determined by the degree of linguistic information that they can encode and whether or not the rules can be structured into a grammar cascade (Appelt et al., 1993) where the output of one rule can be referenced by another rule. The most brittle systems lack a formal grammar and are limited to shallow patterns over sequences of characters. Other systems can incorporate shallow or token-level analysis of the terms in text, such as part-of-speech or phrasal chunk tags (Riloff, 1996; Califf and Mooney, 2003; Drozdzynski et al., 2004; Chiticariu et al., 2010; Chang and Manning, 2014). The most linguistically expressive rule-based systems allow for patterns of syntactic structures (Cunningham et al., 2002; Levy and Andrew, 2006; Valenzuela-Escárcega et al., 2016).

[5] This is not to say such systems are no longer used. Such approaches, while underrepresented in the academic world, are highly valued in industry applications for their great flexibilty, interpretability, and precision (Chiticariu et al., 2013). They may also be used to rapidly annotate documents (Klügl et al., 2016) in a manner that allows for quick iteration and refinement.

Despite possessing a high degree of control and interpretability, rule-based information extraction is often criticized as being far too labor-intensive to be practical (Yakushiji et al., 2006), and having low coverage. This shortcoming has spurred the development of numerous rule learning frameworks and hybridized models designed to extend converage without sacrificing interpretability or customization.

Hybrid approaches A number of rule-based frameworks support rule induction driven by corpus statistics (Riloff, 1996; Huang et al., 2004a; Etzioni et al., 2004), selective application based on observed performance (Soderland, 1999), or incremental generalizations (Califf and Mooney, 1997). Frequently, rule learning approaches operate in a semi-supervised setting, such as bootstrapping (Hearst, 1992; Riloff and Wiebe, 2003) or active learning (Huffman, 1996). Rules may also be used to constrain predictions of statistical models.

While traditional information extraction is limited to a narrow scope, open information extraction aims to extract facts from text in a domain independent manner. There is a substantial body of work addressing open-domain machine reading (Banko et al., 2007; Carlson et al., 2010; Zhang, 2015), as well as systems that target specific domains (Björne et al., 2009; Nédellec et al., 2013; Peters et al., 2014). As Berant et al. (2014) and others have noted, the vast majority of machine reading systems targeting scientific texts focus on reading individual facts or on filling known gaps in a knowledge base.In the biomedical domain, attempts have been made to assemble relations extracted by machine reading into coherent models [6]. This domain is known for the complexity of its models [7], which has spurred research on improved visualizations for actionable insights [8]. This work builds on these previous efforts by assembling complex influence graphs from the extractions of a machine reading system, and through a novel search engine that efficiently searches this graph and visualizes the results in a simple and intuitive interface.

[6] Hahn-Powell, G., D. Bell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). This before That: Causal Precedence in the Biomedical Domain. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pp. 146–155. Association for Computational Linguistics. DOI: 10.18653/v1/W16-2920

[7] Lander, A. D. (2010). The edges of understanding. *BMC Biology*, **8**(1), p. 40. ISSN 1741-7007. DOI: 10.1186/1741-7007-8-40

[8] Dang, T. N., P. Murray, and A. G. Forbes (2015). PathwayMatrix: visualizing binary relationships between proteins in biological pathways. *BMC Proceedings*, **9**(6), p. S3. ISSN 1753-6561. DOI: 10.1186/1753-6561-9-S6-S3

4.4 *Approach*

The approach consists of two stages: (a) machine reading and assembly (MRA), which produces a graph of influence relations from a collection of publications, and (b) a search engine that explores both direct and indirect connections in this graph. This section describes the machine reading and assembly approach. Section 4.5 outlines a search engine over this influence graph.

The methods for machine reading and assembly will cover two scenarios: (a) a domain-specific machine reading system for molecular biology, and (2) an open-domain system for discovering factors influencing children's health. Both systems are rule-based, which has the desirable property of producing an interpretable extraction model that allows for incremental, isolated improvements by an end user that understands the task at hand but is not a natural language processing (NLP), or a machine learning (ML) expert (the likely maintainer of such a system). Although the IE models discussed in the next section are rule-based, the modular design of the system means that the graph explorer interface is agnostic of and independent to the IE component.

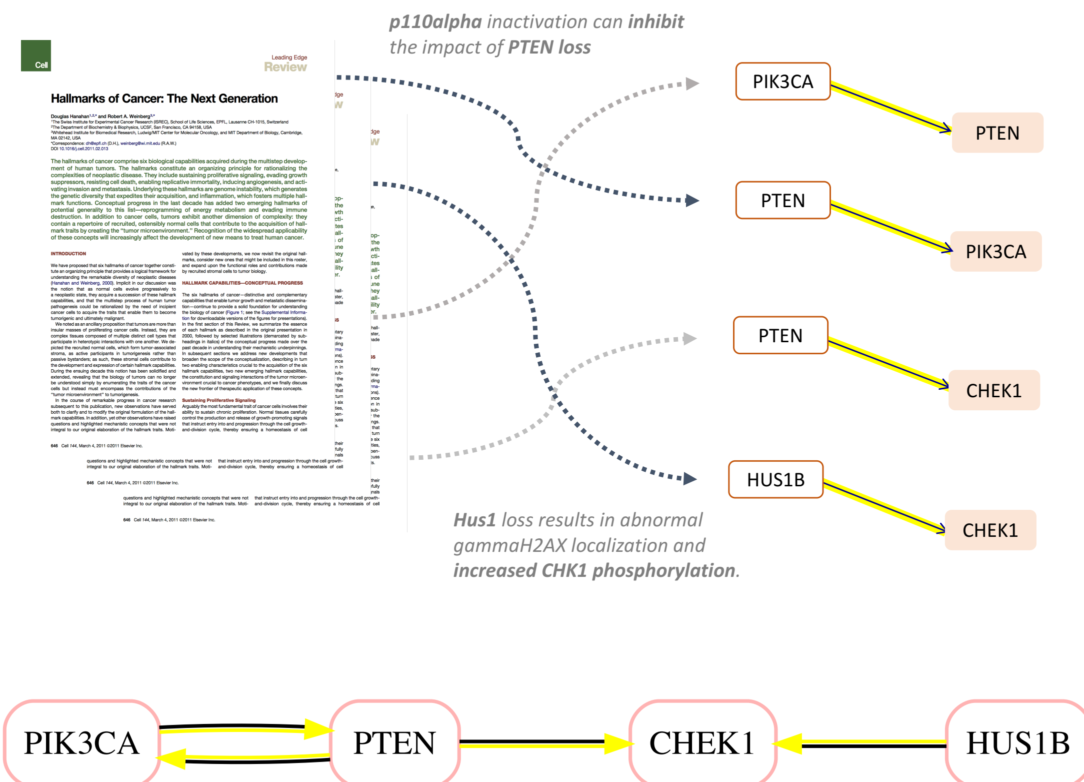


Figure 4.2: Influence relation fragments are extracted from 90,000 scientific publications within the Open Access subset of PubMed.

Domain-specific reading and assembly

Our MRA system for the biomedical literature is called Reach (from REading and Assembling Contextual and Holistic mechanisms from text)[9]. This system extracts entities (e.g., proteins, other chemicals, biological processes) and events (e.g., biochemical interactions) from biomolecular literature. Reach is built on top of the Odin information extraction framework [10] and captures 17 kinds of events, including *nested* events (i.e., events involving other events). The event grammars are applied in cascades composed of rules that describe

[9] `https://github.com/clulab/reach`

[10] Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2016). Odin's Runes: A Rule Language for Information Extraction. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC

patterns over both syntactic dependencies and token sequences, using constraints over a token's attributes (part-of-speech tag, lemma, etc.). The system architecture is summarized in Figure 2.5 of Chapter 2. Reach's assembly component is detailed in Chapter 3 and Appendix A.

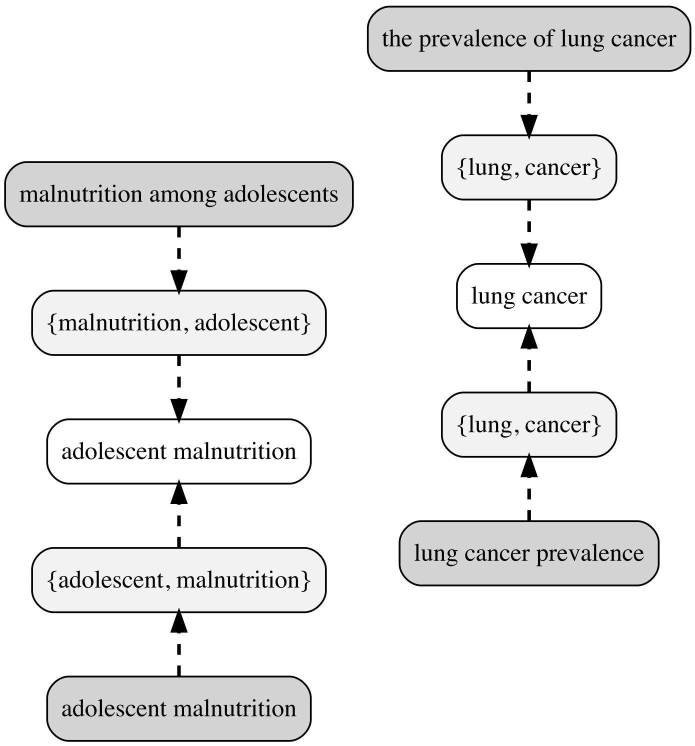


Figure 4.3: Concept mentions are simplified and linked by retaining only the lemma forms of the essential tokens (nouns, adjectives, and verbs).

Multi-domain reading and assembly

Our second use case for MRA models children's health, which involves complex influence chains that span multiple levels of abstraction, linking low-level biomolecular processes with nutritional and socio-economic issues.

In such a setting, comprehensive resources are unavailable. Unlike in Reach, we cannot rely on a taxonomy to guide extractions. While incomplete resources are available for a subset of the covered domains, we treat this use case as an opportunity to explore what can be extracted and assembled when no knowledge base is available for entity recognition or resolution.

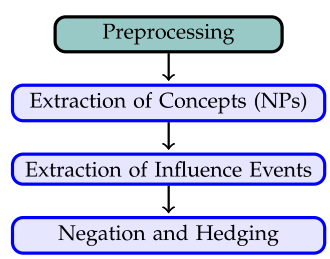


Figure 4.4: Rules for influence events are adapted from Reach grammars pertaining to Activations and Regulations. Entities (concepts) are expanded noun phrases (NPs) in the style of Open IE (Banko et al., 2007).

Entity extraction Following Banko et al. (2007), we instead consider expanded noun phrases as a coarse approximation of the concepts we wish to link. Starting at each noun, we traverse `amod`, `advmod`, `ccmod`, `dobj`, `nn`, `vmod`, and `prep_*` Stanford collapsed dependency relations and promote only the longest span to our pool of candidate concepts. For example, starting at *infections* in "viral infections among infants are [...]", the expansion procedure produces *viral infections among infants* as a candidate entity.

Event extraction and resolution For event extraction, we adapted the Reach grammars that capture influence statements (e.g., positive and negative regulations) to the current task by removing selectional restrictions on the arguments of each event predicate. That is, we extract any lexicalized variation of "A causes B" where A and B are entities identified in the previous step. Matches to these rules produce a directed influence relation that maintains polarity (i.e., increase or decrease). Any mention that is detected as being negated (e.g., "X was not found to increase Y") is discarded from further consideration.

Surviving edges are consolidated through a conservative deduplication procedure where two edges are considered identical if and only if the set of lemmatized content terms[11] for corresponding source-destination concepts between two edges is identical (e.g., "children's stunting" = "stunted children"). When merging duplicates, textual provenance of the extractions is aggregated.

[11] Function words are ignored.

4.5 *Influence graph search engine*

Conventional visualization of any sufficiently complex network suffers from the "hairball" [12] problem. To mitigate this obfuscation, we introduce a search user interface (UI)[13] that allows for structured queries where a user can explore the influence neighborhood around a CAUSE and/or EFFECT to a configurable distance in terms of "hops" in the graph. Alternatively, the user may choose to explore direct and indirect chains of influence linking a possible CAUSE and EFFECT pair (see Figure 4.5 for a detailed description of the UI).

[12] Lander, A. D. (2010). The edges of understanding. *BMC Biology*, **8**(1), p. 40. ISSN 1741-7007. DOI: 10.1186/1741-7007-8-40

[13] `http://influence.clulab.org`

As the pool of analyzed documents grows, possible connections may become so numerous that the results of a query could overwhelm a user. For this reason, we rank query results using a relevance score designed to bring surprising findings to the attention of the user. The scoring procedure is discussed in Section 4.5.

A further means of sifting through the results is provided by a context filter, which allows users to focus results by context, e.g., "show results just for *pancreatic cancer*", or "show impact factors to

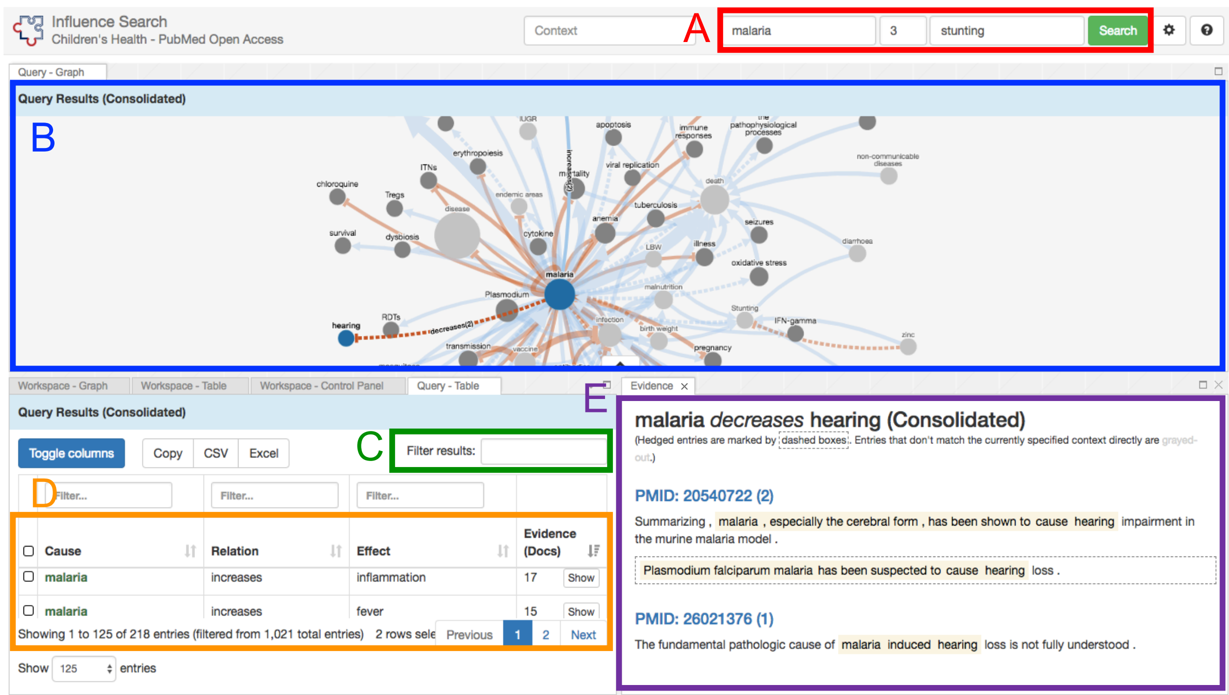


Figure 4.5: The search engine UI for the conceptual influence graph. **A**: search boxes used to specify the purported CAUSE and/or EFFECT, and the maximum number of intervening edges that connect these concepts. Additionally, a context filter can be used to restrict matching edges to those which discuss the term of interest in the papers which provide the provenance for each influence relation. **B**: search box used to further filter query results. **C**: network representation of the query results, displaying how the concepts influence each other. Green edges indicate promotion; red ones indicate inhibition. The width of a link is proportional to the amount of evidence supporting it. **D**: results as a table in which each row corresponds to an edge in the graph shown in **C**. The provenance (textual mentions) of an edge can be viewed by clicking on the *Show* button under the *Evidence (Docs)* corresponding to that row.

children's health in *Sudan*.". The filter constrains results to those influence relations which have at least one piece of evidence originating from a papers that matches the concept term [14]. The presence of a context query affects the display of evidence for each influence relation. Because an influence relation may be supported by evidence from multiple papers, we codify statements visually to mark those originating from papers that meet the specified context constraint.

[14] Term inclusion within a paper can be coincidental, such as in the discussion of related work. To avoid spurious matches and ensure that contextual matches remain topical, we have explored limiting contextual searches to the paper's abstract where any match is likely to prove highly relevant.

Estimating event relevance

In order to rank the results of extraction by an estimate of their relative novelty, each deduplicated edge is scored according to a relevance metric based on the inverse document frequency (IDF) of the lemmatized terms in its concept nodes. Several scores are provided for each edge. These scores differ by (a) whether or not the score incorporates all of the terms in the source and destination concepts or only their head lemmas, and (b) whether the score is an average or maximum. We calculated IDF scores for the lemma of each term in the vocabulary using the entire open access subset of PubMed. To simplify ranking, the scores were normalized using the maximum IDF possible for the dataset.

Score Name	**Description**
AVG IDF	Average normalized IDF for all terms in the source and destination concepts
AVG HEAD IDF	Average normalized IDF of head lemmas for both source and destination nodes in each evidence instance
MAX IDF	Maximum normalized IDF of all terms in the source and destination concepts
MAX HEAD IDF (*default ranking score*)	Maximum normalized IDF of head lemmas for both source and destination nodes in each evidence instance

Table 4.1: A summary of the relevance score metrics calculated for each deduplicated edge. We caclulated IDF scores for the lemma of each term in the vocabulary using the entire open access subset of PubMed. To simplify relevance ranking, the scores were normalized using the maximum IDF possible for the dataset.

Figure 4.5 is an example of output for the domain-specific machine reader (see §4.4). The output demonstrates the system's ability to capture chains of biomolecular interactions. Output for the mixed domain machine reader (see §4.4) is provided in Figure 4.5, which showcases the system's ability to uncover indirect links between

Head-finding procedure

For each piece of textual evidence (mention) for a node,

- Find the highest token leading to a token in the mention's span by following the incoming edges for each token in the mention.
  - Ignore heads that are punctuation.
  - If no head is found for the span, return the last (rightmost) token in the mention's span.

Figure 4.6: The head finding procedure used to generate a subset of the relevance scores.

concepts across domains.

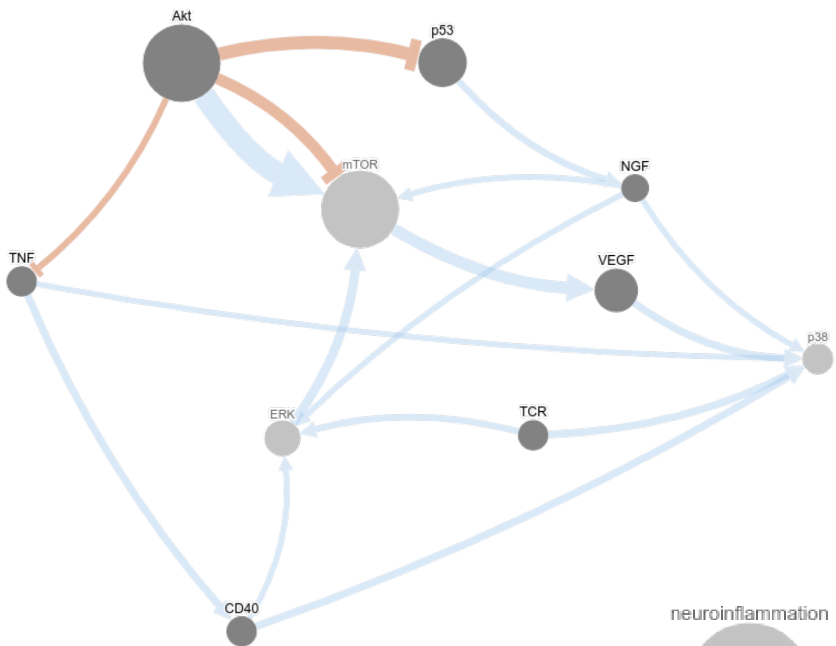


Figure 4.7: An incomplete depiction of *Akt*'s effect on *p38*, which demonstrates multiple pathways.

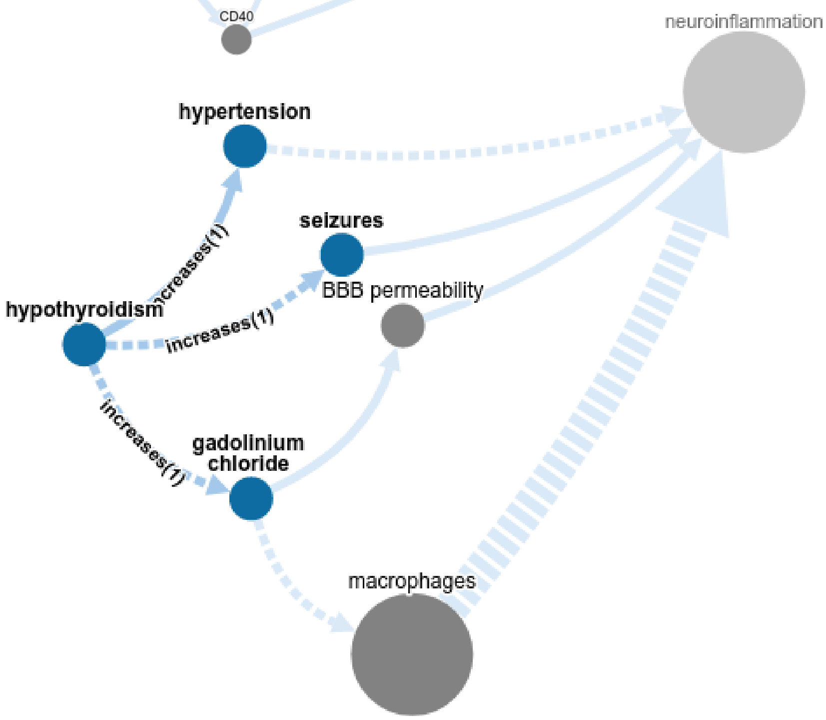


Figure 4.8: *hypothyroidism*'s indirect role in *neuroinflammation*, which demonstrates multiple pathways with varying degrees of confidence. The dotted edges indicate that the majority of the relation's evidence is hedged (ex. *it is suggested that …*).

4.6 *Evaluation*

An early version[15] of this system was evaluated by a biologist who used the tool to augment her model building process for the topic of children's health. The biologist provided the system a set of terms as input. The terms were used to formulate a information retrieval

[15] This was a text-only version.

query against the open access subset of PubMed to identify papers relevant to her use case. The MRA system was then used to extract influence relations from the retrieved documents and present the ranked results to the biologist.

Through her exploration, the biologist refined her search terms of interest and the process was repeated for three cycles over two days. The result was a model containing 35 core concepts such as "EBF", "improved water access", and "government subsidy", and 48 directed influence relations holding between these concepts (e.g., *rapid urbanization* PROMOTES (availability of) *cheap processed food*). Importantly, all influence relations incorporated in her model were substantiated by evidence from the literature.

According to the biologist, the system reduced the model building process from months to two days. She summarized her thoughts in the following quote:

> Finding one of these links manually may well take 1–3 days – sometimes one is lucky and comes across a review paper where someone has already drawn a fairly mature mental model – but filling in the details and making sure they are not biased can take much longer.

This can be thought of as a first step toward bridging research islands through Swanson linking.

4.7 *Conclusions and limitations*

This chapter introduced a search engine that operates over influence graphs that were automatically mined from scientific literature. The approach consists of two high-level components. The first component is a machine reading and assembly system that extracts entities of interest and influence relations that hold between them (e.g., "promotes" or "inhibits"). The reading system can operate in both domain-specific settings, where there are knowledge bases that can support reading (e.g., lists of protein names), and in multi-domain settings where such resources are not available. The second component of our approach is a search engine interface that allows the user to explore not by keywords, but by direct and indirect influence patterns. The results are displayed in a network graph format depicting the subgraph matching the query, and in a table-based view that lists the matching influence relations in descending order of relevance. Through a user study we demonstrated that this deep reading and search approach reduces the effort needed to uncover "undiscovered public knowledge" [16]. With the aid of this tool, a domain expert reduced her model building time from months to two days.

[16] Swanson, D. R. (1986b). Undiscovered public knowledge. *The Library Quarterly: Information, Community, Policy*, **56**(2), pp. 103–118

Improving node and edge deduplication remains a challenging problem in the multi-domain setting where comprehensive knowledge bases are not available. For example, minimal pairs differing in a single modifier such as "*acute* diarrhea" vs. "*chronic* diarrhea" have clinical definitions with clear distinctions. Care must be taken in determining which syntactic constituents (e.g., prepositional phrases) can be safely ignored during comparisons.

Because the reading approach uses a limited set of rules adapted from the domain specific system described in Chapter 2, it is likely this extraction model misses influence statements that may exist in mixed-domain setting. Chapter 5 presents methods for inferring patterns, generalizing rules, and selectively sharing parameterized argument patterns across events involving the same semantic role. Finally, while the multi-domain setting described here may link ideas from different domains, it does so in a way that is oblivous to research communities. Using the influence graph constructed in the multi-domain setting, Chapter 6 explores approaches to predicting future discoveries, and demonstrates how a large citation network can be used to generate semantic representations of research communities. By aligning influence relations with a citation network, we can surface unfamiliar ideas in chains of influence relations which cross distant community boundaries.

5 Learning extraction patterns from few annotations

This idea that there is generality in the specific is of far-reaching importance.

DOUGLAS HOFSTADTER

5.1 Introduction

Despite possessing a high degree of control and interpretability, rule-based information extraction is often criticized as being far too labor-intensive to be practical (Yakushiji et al., 2006). This work presents a simple, automated procedure for deriving an event extraction grammar from annotated data describing biochemical reactions and transferring representative syntactic structures (i.e., the shortest syntactic dependency paths connecting predicate-argument pairs) to detect other event types. The extracted grammar encodes patterns over syntactic structures for better generalizability, and imposes constraints on event arguments to ensure accuracy. Drawing on previous work [1], the grammar is augmented with performance-based thresholding and extended by deriving parameterized templates that can be applied to other event types through rule transfer to improve precision and broaden coverage. These improvements contribute an absolute gain of over 17 points F1 compared to the initial grammar and establish a score well above the median for the 2009 BioNLP Event Extraction Shared Task (Kim et al., 2009). Section 5.2 provides a brief background on event extraction, followed by an overview of the task and an outline of the grammar extraction procedure. Results on the task are discussed in section 5.5. Finally, limitations of rule transfer and performance in low resource settings where annotations are unavailable or incomplete are explored in Sections 5.6 and 5.7.

[1] Riloff, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pp. 1044–1049. AAAI Press. ISBN 0-262-51091-X

5.2 Rule-based event extraction

Event extraction is a technique for automatic reading and summarization that can be used to identify the interaction of entities and the relations between these interactions. Both stochastic and rule-based event extraction systems exist. Though both have demonstrated state-of-the-art results for biomedical event extraction (Hunter et al., 2008; Huang et al., 2004b), the performance ceiling of a stochastic system is often limited by the amount of available annotated training data. In contrast, rule-based systems can be built by a domain expert when training data is unavailable. Rule-based systems have the added advantage of being easily modifiable and transparent to the user by producing a readily interpretable model. Unlike stochastic models, rule-based systems can provide a detailed, step-by-step explanation of how output (i.e., the event representation) is produced from a given document. Historically, rule-based event extraction struggled to avoid explosive rule proliferation to account for variation in text. The current work uses the Odin event extraction framework (Valenzuela-Escárcega et al., 2015b, 2016) to describe an event's underlying syntactic structure as a generalization over textual variation.

Though rule-based systems have been criticized as being labor-intensive endeavors (Yakushiji et al., 2006), this work demonstrates that pattern-based rules for identifying textual mentions of events can be generated automatically by capturing the syntactic structure of event annotations. When annotations are unavailable or incomplete, these structural templates can then be extended across events to broaden grammar coverage and significantly improve recall. This approach proved effective in the 2015 DARPA Big Mechanism (Cohen, 2015; Valenzuela-Escárcega et al., 2017) evaluations on automatic reading [2], where the events being modeled exhibited a high degree of syntactic overlap.

[2] This evaluation is described in detail in Section 2.6 of Chapter 2.

Anatomy of an Odin rule

I will briefly describe the elements of an Odin rule and the steps involved in a rule's application. For a more detailed discussion, I refer the reader to the Odin manual[3] (Valenzuela-Escárcega et al., 2015c) and the description of the rule language given in Valenzuela-Escárcega et al. (2016). Odin supports quantification of patterns (scoped repetitions) to handle the recursive property of natural language. For example, consider the sentence in Figure 5.2 as a case of center embedding:

[3] `http://arxiv.org/abs/1509.07513`

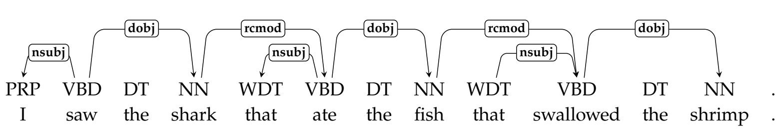


Figure 5.1: A case of centering embedding. The abbreviated set of relations shown in this parse are Stanford collapsed dependencies (de Marneffe and Manning, 2008). `rcmod` refers to a relative clause modifier.

If we wanted to discover the creature at the bottom of this food chain, we could write a rule like that in Figure 5.2 which defines a traversal over the dependency graph. First, though, we identify any simple or compound nouns using the rule shown in Figure 5.2.

```
# comments are preceded by #
# Odin rules are written in YAML
- name: "nouns"
  # the label assigned to the span of text
  # matched by the rule
  label: Noun
  # this rule should run on the first "loop" of the system
  # Odin can run in several iterations or "loops", where
  # later loops make use of the output of previous loops
  priority: 1
  #
  type: token
  # match 1 or more tokens with part-of-speech tags beginning
  # with N (a noun of some kind)
  # a token can be described using any combination
  # of associated attributes (lemma form, tag, etc)
  pattern: [tag=/^N/]+
```

Figure 5.2: A rule finding compound noun NPs as a regular expression over one or more of part-of-speech tags where each tag must begin with `N`.

These simplified NPs will serve as candidate endpoints in the dependency graph traversal defined in Figure 5.2:

```
- name: foodchain
  label: Meal
  priority: 2
  pattern: |
   trigger = [lemma=/^(see|watch)$/]
   apexPredator:Noun = dobj
   foodstuff:Noun+ = dobj rcmod [lemma=/^eat|swallow$/] (rcmod? dobj)+
```

Figure 5.3: A pattern describing a traversal over a syntactic dependency graph. The pattern includes a repeating sequence that ultimately leads to the deepest (lowest) embedded clause.

5.3 *Unsupervised rule transfer*

Rule-based information extraction is often criticized as being impractical due to the associated labor cost. This work presents a simple, automated procedure for deriving an event extraction grammar from annotated data describing biochemical reactions. The grammar encodes patterns over syntactic structures for better generalizability, and imposes constraints on event arguments to ensure accuracy. The grammar is augmented with performance-based thresholding and extended by deriving parameterized templates that can be applied to other event types through rule transfer to improve precision and broaden coverage. These improvements contribute an absolute gain of over 17 points F1 compared to the initial grammar, and establish a score well above the median for the 2009 BioNLP Event Extraction Shared Task.

The majority of events in the biomedical domain are biochemical reactions involving minimally a 'patient'. Hundreds of so-called post-translational modifications (PTMs), an event class which describes enzymatic changes to proteins, have been identified (Pyysalo et al., 2011), though there are annotated examples of only a handful of these categories. Pattern-based rules for identifying textual mentions of such events can be generated automatically from event annotations by extracting the shortest syntactic dependency path between known predicate-argument pairs. Unfortunately, annotated data is limited. This work examines the effects of applying rules generated for one event to a set of related events for which annotations are unavailable or incomplete.

5.4 *The task*

The current work investigates the efficacy of rule thresholding and transfer on improving an automatically extracted grammar for the biomedical domain. The systems are evaluated using the BioNLP 2009 Shared Task on core event extraction[4]. This task is concerned with the extraction of textual mentions in the domain literature of 'bio-molecular events involving proteins or genes in which a participant undergoes a change of state. The task is divided into three subtasks:

[4] `http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/index.shtml#task`

1. Detect event triggers (predicates)
2. Assign event type labels (ex. *phosphorylation* vs. *gene expression*)
3. Identify the core arguments of each event mention (ex. `theme`)

Annotations for the entities[5] are provided. Although events may take other events as arguments, the scope of the current work is limited to examining only those events involving entities as participants. All events that meet this restriction only contain the core argument of THEME, as the task considers events with a CAUSE to be **Regulations** which allow both entities and other events to fulfill the role of CAUSE.

[5] Mentions that can serve as event arguments.

The task makes no distinction between THEME and PATIENT, considering them both to be types of a THEME. To better match the task, the examples that follow assume the same definition.

The event annotations in the training data come from a subset the GENIA event corpus (Pyysalo et al., 2011). In total, the GENIA event corpus covers more than 36,000 event annotations comprised of 36 named categories and an additional category for unknown events; however, this corpus covers only 1,000 scientific paper *abstracts* (Kim et al., 2008).

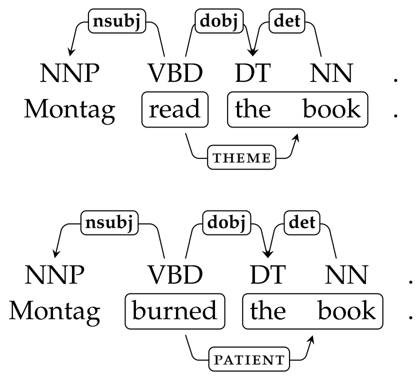


Figure 5.4: Though both sentences are syntactically identical with predicates matching in valency and grammatical features, the thematic role assigned to **the book** differs (THEME in sentence above and PATIENT in the sentence below), due to the lexical semantics of the two predicates.

Rule extraction and domain transfer

The current work uses the training corpus annotations to derive event rules. For each event annotation, the system extracts the shortest syntactic dependency path between a predicate ("trigger") and its THEME. It then generates two rules: (1) the event-specific form and (2) a templatic version, which allows the argument pattern to be paired with other event predicates (see Figure 5.4 for an example). The performance baseline is then simply to use the event-specific form (i.e., only apply the event patterns observed in the training data for each event type). For each event type, the system extracts a set of rules in training and then apply them to the development partition of the task data.

An examination of the training data for the task reveals an unbalanced distribution of event types. Several event types are relatively uncommon in the training data (see Table 5.4).

Argument	Frequency
THEME	886
THEME_2	243
THEME_3	5
THEME_4	2

Table 5.1: Frequency of various **Binding** arguments. **Binding** is the only event type in this task which allows for multiple THEMES.

Distribution of events in training data

Event	Frequency
Gene Expression	1770
Binding*	1136
Transcription	631
Localization	258
Phosphorylation	185
Protein Catabolism	138

Table 5.2: Distribution of THEME instances across each event type in the training data (excluding regulations). **Binding** events involve multiple THEMES with varied and largely inconsistent syntactic context. **Gene Expression** event mentions dominate the training data, being an order of magnitude more common than mentions of **Protein Catabolism**.

Experimentally, an attempt to improve recall is made by sharing observed syntactic positions of a THEME across event types. Argument patterns extracted for one event are modified so that they can be used by a different event. Minimally, this involves replacing an observed predicate with a disjunction of all predicates known to be associated with the target event. Essentially, rules extracted for each argument type are then consolidated and transformed into *templatic* rules that may be used by any event involving the target argument type.

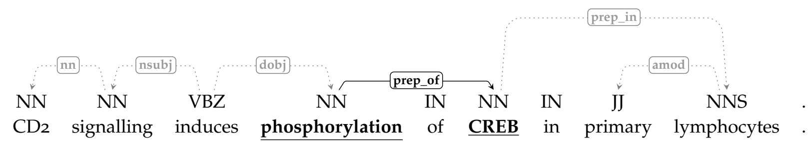


Figure 5.5: A sentence containing a **Phosphorylation** event. The event trigger, *phosphorylation*, is a nominal predicate connected to the event THEME, **CREB**, through an outgoing `prep_of` (preposition *of*) syntactic dependency relation.

The THEME of the **Phosphorylation** event shown in Figure 5.4 can be encoded as a syntactic pattern using the Odin[6] rule shown in Figure 5.4.

[6] For details on the Odin rule language, see Valenzuela-Escárcega et al. (2016)

Extracted rule

```
label: Phosphorylation
pattern: |
  trigger = [lemma=phosphorylation]
  theme:Protein = >prep_of
```

Figure 5.6: A rule extracted from an event annotation in the training data. The rule describes a **Phosphorylation** event with a THEME as the head of an of-PP complement attached to the nominal predicate *phosphorylation*.

We can then transfer knowledge of the observed syntactic position of one event's argument to the same argument (THEME) in underrepresented events through rule adaptation and parameterization (see Figures 5.4 and 5.4).

Templatic rule

```
label: ${EventLabel}
pattern: |
  # templatic representation (using variables)
  trigger = ${nominalPredicate}
  theme:Protein = >prep_of
```

Figure 5.7: The pattern in the previous example is adapted to capture any event, given the event's label and the pattern describing that event's nominal predicates. Here `${nominalPredicate}` refers to a variable that expects a disjunction of nominal predicates.

Adapted rule

```
label: Gene_expression
pattern: |
  trigger = [lemma=expression]
  theme:Protein = >prep_of
```

Figure 5.8: The pattern in the previous example is adapted to capture a **Gene Expression** event. Creating the **Gene Expression** in this context means only changing its label and predicate pattern (`${nominalPredicate}` → `[lemma=expression]`). This is made more general by allowing any nominal predicate that has been observed in a Gene Expression event mention.

Path lexicalization

To maximize rule generalizability, wholly *unlexicalized* paths for event arguments are used[7]; each "hop" in the predicate→argument path is unconstrained by features related to the identity of its tokens. Using the sentence and accompanying syntactic dependency parse presented in Figure 5.9, a **lexicalized** form of the relational dependency path originating from the predicate **ate** and terminating at "Ayu" is provided in Figure 5.10. The **unlexicalized** form of the same path given is provided in Figure 5.11. Unlexicalized paths are readily transferrable to other event types, as they are largely unconstrained by event-specific information[8].

[7] In this work, we use the Stanford collapsed dependencies (de Marneffe and Manning, 2008), which handles prepositional phrases by including the preposition in the edge (relation)'s label. For example, "phosphorylation of A" is encoded as the vertex *phosphorylation* $\xrightarrow{\texttt{prep\_of}}$ *A*. This representational feature means that even "unlexicalized" paths involving traversals of a prepositional phrase (PP) will include a reference to the specific preposition heading the PP.

[8] See Figures 5.4, 5.4, and 5.4.

5.5 *Results*

Rules were extracted from the training data using the procedure described in Section 5.4. The effect of rule transfer and performance-based thresholding of rules was compared to a baseline system using neither rule transfer nor thresholding.

Baseline

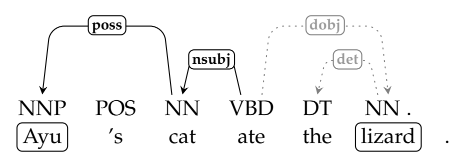


Figure 5.9: A simple sentence with a nominal subject selected to illustrate varying degrees of lexicalization possible in a relational dependency path.

```
>nsubj [lemma=cat] >poss [tag=NNP]
```

Figure 5.10: A lexicalized path from "ate" → "Ayu" in the sentence shown in Figure 5.9.

```
>nsubj >poss
```

Figure 5.11: A unlexicalized path from "ate" → "Ayu" in the sentence shown in Figure 5.9.

Performance of the adapted rules on the development partition is compared with a baseline measure where for each event, rules are extracted in the training data and then applied to all instances of that corresponding event in the development data. As Table 5.5 shows, this approach provides a relatively strong baseline with generally high recall.

An examination of the training data for the task reveals an unbalanced distribution of event types (see Table 5.4). The experiments described in Section 5.5 test whether this underrepresented set of events can benefit from a form of transfer learning or domain adaptation (Thrun and O'Sullivan, 1995; Silver, 1996) where patterns of one event (the source domain) type are selectively transferred or adapted to another event (the target domain) type.

It is worth noting that **Binding** events involve multiple "themes" (ex. Theme$_2$, Theme$_3$, Theme$_4$), though these arguments may appear in canonically agentive positions:

Results of baseline approach

Extract from → Apply to	Gold	Predicted	Matched	R	P	F
Gene Expression	356	439	247	69.38	56.26	62.14
Transcription	82	418	47	57.32	11.24	18.80
Protein Catabolism	21	19	15	71.43	78.95	75.00
Phosphorylation	47	40	30	63.83	75.00	68.97
Localization	53	233	31	58.49	13.30	21.68
Simple Event Total	559	1149	370	66.19	32.20	43.33
Binding	248	239	52	20.97	21.76	21.36
Event Total	807	1388	422	52.29	30.40	38.45

Table 5.3: Rules extracted from the training data were applied to the corresponding event in the development set. The complete triggers were used as-is, without any simplification. *Gold* refers the the total number of true instances of each class. *Predicted* refers to the number of mentions predicted for each class, and *Matched* refers to the number of the correct predictions (true positives) for each event class. *R* corresponds to recall (the percentage of true instances the system was able to produce). *P* refers to precision (the number of correct predictions for each class over the total number of predictions made for that class). *F* refers to the F1 or harmonic mean of precision and recall.

1. **A** binds with *B*
   - **A** appears in a VP-external position.
2. **B** is bound by *A*
   - Valency-decreasing passivization where agentive **A** appears as a *by*-PP adjunct.

Experimental

In the experimental setting, we explored a number of configurations for path lexicalization and performance-based thresholding. Here only results for unlexicalized paths are reported to highlight the effects of rule transfer across events, as fully lexicalized paths showed little to no transfer[9]. The experiments measure the contribution of a rule template constructed from the cross-event union of all distinct paths observed in the training data beginning at an event predicate and leading to the event's THEME. The application of this template to each event class was constrained to matches of a predicate pattern rule representing the disjunction of lemmatized predicate sequences observed for the event in the training data. As noted earlier, **Binding** events contain multiple themes, some of which appear syntactically as agents. Due to their syntactically ambiguous nature, patterns derived from binding events were not applied to other events. Adapting or consolidating THEME_2 - THEME_4 rules for transfer to events requiring exactly one THEME is considered a task-specific problem and outside the scope of the current work[10].

Feature selection of rules was performed by tuning performance thresholds on the training data (see Table 5.5). The results reported are on the development data, which was reserved for testing. Notice that precision for **Phosphorylation** is largely unaffected by the

[9] Fully lexicalized paths are generally too constrained to transfer to other events, as participants and intervening tokens would need to be identical to match.

[10] One obvious approach is to map extracted patterns for both THEME and CAUSE to THEME when applied to **Binding** events, though the inconsistency of the labeling of $\text{THEME}_1 \ldots \text{THEME}_n$ in the task data remains a problem.

inclusion of consolidated event templates (cf. Tables 5.5 & 5.5). This is likely due to the unambiguous nature of **Phosphorylation** event predicates, which exhibit low variety when compared to other events (see Table 5.5).

Results of transfer without performance-based thresholding

Event	Gold	Predicted	Matched	R	P	F
Gene Expression	356	1115	287	80.62 (**+11.24**)	25.74	39.02
Transcription	82	836	64	78.05 (**+20.73**)	7.66	13.94
Protein Catabolism	21	25	15	71.43	60.00	65.22
Phosphorylation	47	48	35	74.47 (**+10.64**)	72.92	73.68
Localization	53	491	42	79.25 (**+20.76**)	8.55	15.44
Simple Event Total	559	2515	443	79.25 (**+13.06**)	17.61	28.82
Binding	248	672	76	30.65 (**+9.68**)	11.31	16.52
Event Total	807	3187	519	64.31 (**+40.72**)	16.28	25.99 (**-12.46**)

Table 5.4: Performance of our rule transfer system on the development set of the BioNLP 2009 Shared Task (see Table 5.5 for a description of the columns). Rules were extracted from the training data and applied to the development data. In addition to the original rules, this grammar includes a consolidated template representing all distinct extraction rules for the argument THEME across event categories. Since no feature selection was performed on the patterns extracted from the training data, this grammar represents the recall ceiling of this approach.

The next set of results show the effect of performance-based thresholding of rules. The fitness of a rule for inclusion in the system can be evaluated by imposing threshold constraints on precision, recall, F1, and frequency values. These thresholds were tuned to maximize performance on the development data. Since any form of transfer should improve recall, the critical evaluation metric to balance performance (F1) on unseen data is precision. This was confirmed empirically through a search over parameter values for precision, recall, F1, and count thresholding. A path's performance may be evaluated per event category or across event categories. Contrary to previous work on domain adaptation, it was found that relying on path scoring per event for thresholding yielded better results than thresholding by path performance across events. The value of per-event performance on building the "general" domain is unexpected and likely due to the high degree of syntactic variation and low overlap in event representations observed across these event

Distinct lemmatized predicate patterns per event

Event	Distinct predicate patterns
Gene Expression	83
Transcription	47
Protein Catabolism	15
Phosphorylation	10
Localization	39

Table 5.5: Counts of the lemmatized predicate (trigger) patterns extracted in the training data. **Phosphorylation** and **Protein Catabolism** event categories exhibited the lowest diversity of triggers. The rule transfer template's application to each event type requires a match of one of these lemmatized sequences.

Results of transfer with performance-based thresholding

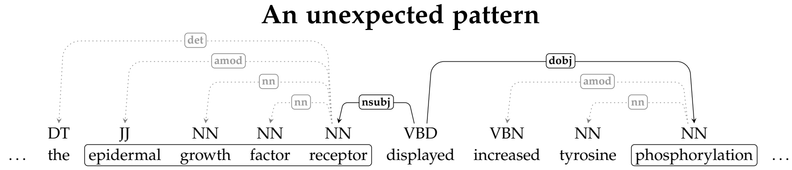
Event	Gold	Predicted	Matched	R	P	F
Gene Expression	356	287	228	64.04	79.44	70.92 (**+8.78**)
Transcription	82	52	31	37.80	59.62	46.27 (**+27.47**)
Protein Catabolism	21	17	14	66.67	82.35	73.68 (**-1.32**)
Phosphorylation	47	40	30	63.83	75.00	68.97
Localization	53	37	31	58.49	83.78	68.89 (**+47.21**)
Simple Event Total	559	433	334	59.75	77.14	67.34 (**+24.01**)
Binding	248	93	38	15.32	40.86	22.29 (**+0.93**)
Event Total	807	526	372	46.10	70.72	55.81 (**+17.36**)

Table 5.6: Results of our top performing experimental system on the development set of the BioNLP 2009 Shared Task (see Table 5.5 for a description of the columns). The system uses both rule transfer across events and precision-based thresholding of rules. The event grammar was extracted from the training data. In addition to the original rules, this grammar includes event-specific templates allowing for the lemmatized forms of any of an event's known predicate (trigger) patterns to satisfy its extracted rules. Each rule of the grammar was then subjected to a precision threshold based on its performance over the training data. A precision threshold of 50% was found to be optimal after tuning.

types (see Section 5.6 for details). Event-specific evaluation of a event-argument path is also arguably a more natural metric in a real-world case where a path's suitability across events is wholly unknown. The effect of precision-based thresholding within each event category is reported in Table 5.5. This grammar includes a consolidated template representing all distinct extraction rules for the argument THEME across event categories. Each rule of the grammar was then subjected to a precision threshold based on its performance on the training data. A precision threshold of 50% was found to be optimal after tuning.

In addition to performance gains, feature selection through precision thresholding of rules yielded a 55% reduction in rules (see Table 5.5). The automated rule extraction procedure when coupled with precision-guided rule transfer draws attention to reasonable paths that are unlikely to be devised manually (see Figure 5.5 for an example).

An unexpected pattern




```
label: Phosphorylation
pattern: |
trigger = [lemma="phosphorylation"]
Theme:Protein = <dobj >nsubj
```

Figure 5.12: A pattern with 57% precision that is well-formed, though unexpected.

Reduction of event-specific rules

Event	Original rules	After thresholding
Gene Expression	888	486 (**-54%**)
Transcription	450	212 (**-47%**)
Protein Catabolism	120	67 (**-56%**)
Phosphorylation	130	86 (**-66%**)
Localization	254	163 (**-64%**)
Binding	1206	803 (**-67%**)
Original total	3048	1817 (**-60%**)
	Distinct templates	
	(before thresholding)	(after thresholding)
	455	126 (**-67%**)
Total	3503	1943 (**-55%**)

Table 5.7: Summary of rule reduction effects of precision-based thresholding. The count of original rules involves duplicate argument patterns in cases where an argument pattern was observed with different predicates.

Distinct argument patterns

Event	**Distinct argument patterns** (original with thresholding)	**+ Templates**
Localization	23	135
Phosphorylation	29	140
Transcription	32	140
Binding	202	225
Gene Expression	57	153
Protein Catabolism	10	130

Table 5.8: Additional argument patterns through template inclusion. Precision-based thresholding of templates is a conservative effort at increasing recall. Notice the counts given here are lower than those in the *original rules* column of Table 5.5, as the original rules may involve repetitions of an argument pattern with different predicates.

Comparison to other systems

Table 5.5 shows the top-performing system for the 2009 task. Though the top-performing system was not rule-based, the results of the current work generally meet or exceed the scores of this system, save for the **Binding** event type which was largely ignored for this task due to irregularities involving multiple themes. Experimental results exceed the median scores reported for the task in Table 5.10. The current system's performance indicates that a rule-based system can in many cases outperform a traditional machine-learning system.

Top-performing system for task

Event	R	P	F
Gene Expression	68.56	77.10	72.58
Transcription	36.50	64.10	46.51
Protein Catabolism	42.86	66.67	52.17
Phosphorylation	76.30	70.55	73.31
Localization	48.85	80.95	60.93
Binding	34.58	43.01	38.34
Event Total	56.18	68.23	61.62

Table 5.9: The top-performing system for the task in 2009 (University of Turku). These results reflect performance on the **test** data.

	Test
Median Recall	37.67
Median Precision	58.74
Median F1	43.45

Table 5.10: Median scores for systems performance on the **test** data.

5.6 *Explaining the limitations of cross-event transfer*

Our results indicate that performance-based thresholding is more effective than rule transfer for the events examined. Combining these two approaches provides no noticeable improvement over the event-specific rules when thresholded for performance. Low resource simulations described in Section 5.7, however, suggest that this is due in large part to the size of the training data being sufficient to cover the majority of syntactic contexts for each event. To understand the apparent ineffectiveness of rule transfer across events, features vectors for each event were constructed using the per-event precision of each element of the set of distinct patterns modeling the syntactic path from an event's predicate to its THEME. These paths comprise the observed syntactic "vocabulary" linking any predicate to its THEME. Precision values for each event were calculated by applying the consolidated template rules over the training data.

We measured the syntactic relatedness of the events as the pairwise cosine similarity of all event pairs. For each event, we produce a vector of precision scores for each argument pattern mined from the training data. Table 5.6 shows the results of the similarity comparison.

Similarity of events by the precision of mined argument patterns

	Protein Catabolism	**Localization**	**Transcription**	**Binding**	**Phosphorylation**	**Gene Expression**
Protein Catabolism	1.00	0.04	0.05	0.17	0.22	0.08
Localization	0.04	1.00	0.03	0.18	0.09	0.04
Transcription	0.05	0.03	1.00	0.05	0.05	0.08
Binding	0.17	0.18	0.05	1.00	0.15	0.19
Phosphorylation	0.22	0.09	0.05	0.15	1.00	0.11
Gene Expression	0.08	0.04	0.08	0.19	0.11	1.00

Table 5.11: Themes across events show a low similarity in terms of their syntactic context. Scores range from 0 (unrelated) to 1 (identical).

Surprisingly, the events are quite distinct in their syntactic representations of the argument THEME. The most similar pair of events, **Phosphorylation** and **Protein Catabolism**, have only a 0.22 similarity score. This in part explains the lack of improvement from transfer; what similarity exists between these events is likely already captured individually.

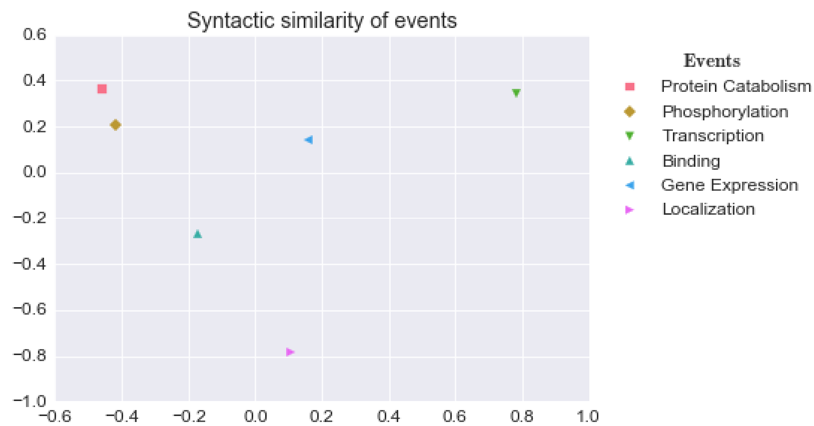


Figure 5.13: Visualization of syntactic similarity of events. For each event, a feature vector was composed by measuring the precision of each path in the set union of syntactic paths leading to a THEME. Cosine similarity scores were then calculated for each pair of events, resulting in a vector of six features for each event (see Table 5.6).

Further analysis reveals that few events demonstrate the linguistic expectations of a THEME, such as a reliable appearance as a direct object in the case of a verbal predicate. Table 5.6 shows the performance of a set of linguistically predictable syntactic configurations for a THEME, as well as an unexpected case where the THEME appears as a nominal subject (`>nsubj`).

Precision of selected syntactic patterns for theme

	`>nsubjpass`	`>dobj`	`>nsubj`	`>nn`	`>prep_of`	`>prep_with`
Gene Expression	54.17	39.88	12.14	53.27	55.92	45.45
Transcription	5.97	5.17	0.0	24.14	8.28	0.0
Protein Catabolism	1.0	66.66	0.0	86.67	80.95	0.0
Phosphorylation	1.0	71.43	0.0	93.75	96.15	0.0
Localization	10.15	18.46	34.48	13.21	13.15	0.0
Binding	29.17	32.18	66.66	35.47	56.69	27.45
Across events	30.86	30.56	25.35	33.96	35.17	28.79

Table 5.12: Precision of selected syntactic paths for each event.

An examination of the top three most common paths with a frequency ≥ 3 and precision $> 40\%$ for each event (Table 5.6) highlights some of the syntactic dissimilarity. The results demonstrate that the syntactic position of an argument is insufficient for disambiguating its thematic role, which is determined largely by the lexical semantics

Most common patterns per event

Event	Path	Mean Precision (%)	Count	Path Length
Binding	`Theme:Protein = >vmod >dobj`	75	4	2
	`Theme:Protein = >nsubj >dep`	71	7	2
	`Theme:Protein = >nsubj`	67	78	1
Gene Expression	`Theme:Protein = >prep_of >conj_negcc`	80	5	2
	`Theme:Protein = >prep_of >prep_including >appos`	80	5	3
	`Theme:Protein = >conj_and >dep`	78	9	2
Localization	`Theme:Protein = >nsubj >appos`	50	6	2
	`Theme:Protein = >prep_of >prep_of >appos`	50	6	3
	`Theme:Protein = >dobj >appos`	33	6	2
Phosphorylation	`Theme:Protein = >nsubjpass`	100	6	1
	`Theme:Protein = >prep_of >dep`	100	6	2
	`Theme:Protein = >prep_of`	96	52	1
Protein Catabolism	`Theme:Protein = >nsubjpass`	100	7	1
	`Theme:Protein = >nn`	87	15	1
	`Theme:Protein = >prep_of`	81	42	1
Transcription	`Theme:Protein = >prep_from >nn`	50	6	2
	`Theme:Protein = >prep_including`	40	5	1
	`Theme:Protein = >dobj >prep_for`	38	8	2

Table 5.13: The top three most common paths for each event with a frequency ≥ 3 and precision $> 40\%$. The overall low precision of Localization and Transcription paths suggests these events have ambiguous triggers and would benefit from greater lexicalization in the representation of their paths. Note also the appearance of `_negcc`, which illustrates that the task seems to consider mentions of the form "... **expression** *of* ... *nor* **...**" to be a positive instance of an event mention, due to its implied existence in another context.

of the predicate. The event with the highest precision in Table 5.6, **Phosphorylation**, possesses a small set of candidate predicates each of which is entirely unambiguous as inflections of the event label (e.g, *phosphorylates*, *phosphorylated*, etc.).

5.7 *Low resource settings*

Though rule transfer has shown to be ineffective with the current task when extracting event-specific rules from the entire training data, in low resource settings where annotated training data is unavailable rule transfer can provide a "free" boost to performance. A low resource setting can be simulated by randomly sampling a small portion of annotations. The random sampling should reflect the relative frequency of events in the training data. Figure 5.7 shows the effect of rule transfer when limited to using a random sampling of 1-5% of the available training data.

Though the events are largely dissimilar when considering the syntactic context of each event across the entire training data, rule transfer improves performance when annotations are sparse by increasing representativity of those syntactic structures common to all events. As the number of annotations increase, we see that transfer

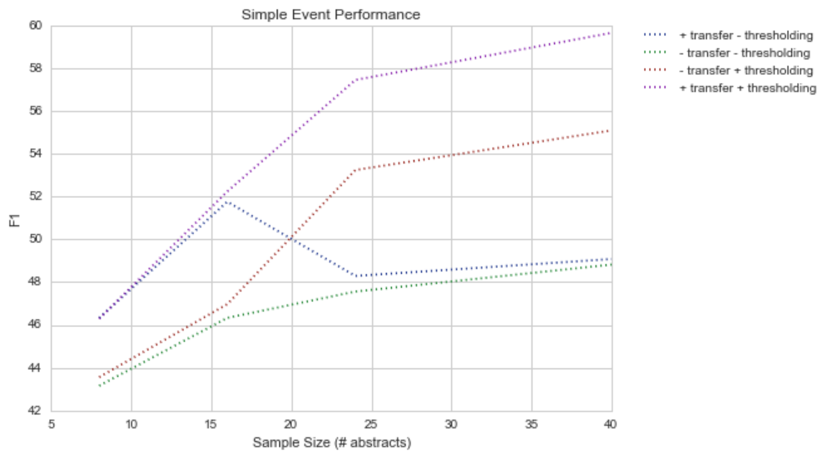


Figure 5.14: Results of the experimental approach in low resource settings on the development data. When considering paths extracted for event annotations of 40 abstracts, a combination of rule transfer with performance-based thresholding yields a 10 point absolute gain to F1 over the baseline and a 5 point gain over thresholding alone.

without thresholding introduces noise.

5.8 *Conclusions and limitations*

This chapter outlined a generalized method for automatically extracting human readable event rules from annotated training data using unlexicalized dependency paths between an event predicate and each of its arguments. The approach can be used to rapidly bootstrap a grammar for a machine reading system like the one described in Chapter 2 whenever annotated data is available. While hundreds of biomolecular events have been reported in the literature, fewer than a dozen are well-represented in annotated corpora (Pyysalo et al., 2011). Rule transfer has the potential to greatly improve coverage of the long tail of underrepresented events. Though further work is needed on labeling events in cases with ambiguous predicates, rule transfer may be used to improve coverage, especially in low resource settings. Coupling pattern induction and transfer approaches with precision-based thresholding results in a large overall improvement to performance. In fact, this manner of rule-based event extraction can outperform machine learning methods for the same task.

6 *Opportunity discovery*

There is no magic. There is only knowledge, more or less hidden.

Gene Wolfe
Shadow & Claw

6.1 *Overview*

This chapter examines opportunities for scientific discovery and collaboration based on the work introduced in earlier chapters. I will begin by outlining a method for predicting future scientific discoveries that we first introduced in Luo et al. (2018)[1]. The approach is built upon the conceptual influence graph described in Chapter 4. The second type of opportunity model we will examine in this chapter is a method for measuring the relatedness or isolation of research communities based on a large citation graph of over 28 million links. By aligning communities to influence relations, we can rank chains of causal statements using an estimate of the involved communities' mutual awareness to reveal pathways which bridge distant communities and surface "undiscovered public knowledge" (Swanson, 1986b). To illustrate this capability and emphasize the generalizability of the approach, I will introduce several examples of novel hypotheses addressing a multitude of afflictions and involving research from seemingly distant communities.

[1] Luo, F., M. A. Valenzuela-Escárcega, G. Hahn-Powell, and M. Surdeanu (2018). Scientific Discovery as Link Prediction in Influence and Citation Graphs. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pp. 1–6. Association for Computational Linguistics. DOI: 10.18653/v1/W18-1701

6.2 *Predicting future discoveries*

In Luo et al. (2018), we introduced a machine learning approach inspired by the field of link prediction (Liben-Nowell and Kleinberg, 2007; Leskovec et al., 2010) for the identification of "white spaces" in scientific knowledge (Sebastian et al., 2017; Cameron, 2015). We leverage an influence graph of cause-effect statements[2] mined from over 100K publications [3] to approximate the task of white space discovery by predicting new influence relations that do not exist in

[2] For example, *smoking increases the risk of colorectal carcinomas* or *fat malabsorption leads to steatorrhoea and malnutrition*.

[3] Because the hypotheses pursued in these publications are generally expressed using causal language, we believe this influence graph captures the essence of the scientific knowledge in this domain.

the graph at a given time t, but are found to emerge at some later time.

Approach

To reduce noise, only influence relations extracted at least twice and linking concepts with an inverse document frequency (IDF) larger than 1 were retained. This filtering criteria drastically reduces the size of the influence graph of Hahn-Powell et al. (2017) by two orders of magnitude from nearly 1.5 million distinct nodes linked by 2.4 million directed influence relations to 30,000 concepts and 69,000 relations. Our pruned influence graph is accompanied by a citation graph containing roughly 5.5 million outgoing citations[4] from the 100,000 source publications.

[4] These citations involve over 2.6 million distinct publications.

To reiterate, the aim of this work is to predict which influence relations will be added to the influence graph after time t using only information available prior to t. We consider this to be a reasonable approximation for the discovery of gaps in science knowledge, since influence relations that will be added in the future indicate that information was deemed important enough to be researched and disseminated through publication. Given the size of the influence graph, we elected to limit the search space considered by narrowing our predictions to cases of A $\rightarrow$ C influence relations, when A $\rightarrow$ B and B $\rightarrow$ C exist in the graph before time t. Figure 6.2 illustrates this constraint. This is a nontrivial formulation of the problem, since transitivity cannot be assumed to be true; influence relations extracted from text frequently oversimplify complex causal processes. Section 6.2 demonstrates that relying on this transitivity assumption leads to poor predictions.

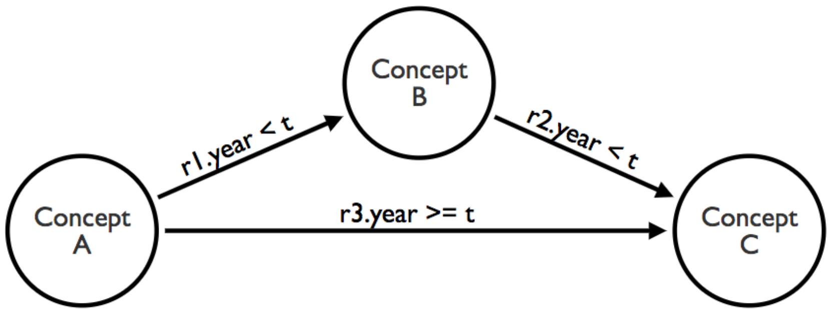
We implemented the prediction task as a form of backtesting where we select a year of publication as a partition point t, and create positive training examples from the A $\rightarrow$ C relations that were added to the influence graph after t. Similarly, we create negative examples from A $\rightarrow$ C links that do not appear between time t and the present. These negative examples are admittedly an approximation in that some of these relations may correspond to discoveries that will be published at some later date following the most recent publication in our dataset. In this work, we chose 2012[5] as our value of t based on an analysis of the distribution of positive and negative examples for different values of t represented in the pruned influence graph (see Table 6.1 for details).

[5] We chose 2012 for our partitioning point t because the split provided a sufficiently high number of positive examples, while still maintaining a realistic distribution of negative examples.

t (year)	positive	negative
2017	-	998,586
2016	319	979,709
2015	1,706	881,689
2014	3,767	696,406
2013	5,208	481,671
2012	**5,015**	**274,251**
2011	3,741	151,460
2010	2,448	73,226
2009	1,521	36,839
2008	782	16,372
2007	444	7,843

Table 6.1: The total number of positive and negative examples for different values of the time threshold t.

Using our selected value for t, the dataset was split into training, development, and testing partitions as indicated in Table 6.2, using a 60-20-20% split. During the partitioning, we took care to ensure that identical influence links coming from different papers were allocated to the same partition.



	Positive Examples	Negative Examples
Training	3,011	164,551
Development	1,002	54,850
Testing	1,002	54,850

Table 6.2: 60-20-20 split of positive and negative examples in the training, development, and testing partitions.

Figure 6.1: We predict if links such as A → C will be added to the influence graph after time *t*, using information that exists before time *t* such as A → B, and B → C. Setting $t = 2012$ yields 5,015 positive examples, i.e., A → C links that were added after 2012, and 274,251 negative examples, i.e., A → C links that were *not* added to the graph between 2012 and the end of 2017.

As shown in Section 6.2, we experimented with a variety of both linear and nonlinear classifiers. A key contribution of this work is the set of features devised for this task. The feature set, summarized in Table 6.2, captures the connectivity of both the influence and citation graphs around a candidate link. The guiding assumption motivating this is that the more connected the corresponding graph is around concepts A and C, the more likely it is that the link $A \rightarrow C$ will be discovered in the near future (in other words, a great deal of attention centers around highly connected concepts). From the influence graph, we extract the in- and out-degrees of the source and destination nodes, along with statistics from the path(s) connecting the two nodes. The statistics include the length of the shortest path linking the source and destination nodes, or the inverse document frequency (IDF) scores of the nodes along such paths.

Based on the idea that discoveries are more likely if the individual fragments that form the puzzle ($A \rightarrow B$ and $B \rightarrow C$ here) come from the same or related discipline, we derive features from the citation graph based on the probabilities that papers containing $A \rightarrow B$ ($\mathbf{p}_{A \rightarrow B}$) and $B \rightarrow C$ ($\mathbf{p}_{B \rightarrow C}$) belong to the same communities. We model the probability that two papers, $p1$ and $p2$, belong to the same community $P(p1, p2)$ using two configurations of the Coda community detection algorithm [6]. The first configuration is parameterized to detect 100 communities, while the second is set for 300. Because influence relation in the pruned graph are reported in a minimum of two publications, we derived the max/min/avg P($p_{A \rightarrow B}$,$p_{B \rightarrow C}$) features, which are computed across all possible combinations of papers $p_{A \rightarrow B}$ and $p_{B \rightarrow C}$.

[6] Yang, J., J. McAuley, and J. Leskovec (2014). Detecting Cohesive and 2-mode Communities Indirected and Undirected Networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pp. 323–332. ACM, New York, NY, USA. ISBN 978-1-4503-2351-2. DOI: 10.1145/2556195.2556243

Finally, we add a series of features extracted from the collection of biomedical publications used in these experiments, such as IDF scores of the relevant concept nodes, and the number of papers that mention a given influence relation.

Features for link prediction models

Feature Name	Description
C_A(outdegree)	Out-degree of source concept node A, i.e., number of influence relations starting on A
C_A(indegree)	In-degree of source concept node A, i.e., number of influence relations ending on A
C_C(outdegree)	Out-degree of destination concept node C
C_C(indegree)	In-degree of destination concept node C
C_{between}(outdegree)	Out-degree of nodes in all the shortest paths that connect A to C but do not pass through B
C_{between}(indegree)	In-degree of nodes in all the shortest paths that connect A to C but do not pass through B
shortest_path_length	The length of the shortest path that connects A to C but does not pass through B; 0 if no such path exists
shortest_path_count	The number of shortest paths that connect A to C but do not pass through B
C_{between}(avg-idf)	Average inverse document frequency (IDF) of nodes in-between A and C in all the above shortest paths
r_{between}(avg-seen)	Average number of papers containing an edge in the above shortest paths
max P($p_{A\to B}$,$p_{B\to C}$)	Maximum probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm; p_r refers to any paper that contains influence relation r.
min P($p_{A\to B}$,$p_{B\to C}$)	Minimum probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm
avg P($p_{A\to B}$,$p_{B\to C}$)	Average probability of papers $p_{A\to B}$ and $p_{B\to C}$ being related based on their membership to multi-communities detected by the Coda algorithm
Jaccard($\mathbf{p}_{A\to B}$,$\mathbf{p}_{B\to C}$)	Jaccard similarity between the set of papers that contain the link $A \to B$ ($\mathbf{p}_{A\to B}$) and the set of papers that contain $B \to C$ ($\mathbf{p}_{B\to C}$)
Inter-citation ratio	The number of citations between the two sets $\mathbf{p}_{A\to B}$ and $\mathbf{p}_{B\to C}$ normalized by the size of the union of the two sets.
C_A(IDF)	IDF of the lemmatized terms of source concept node A
C_B(IDF)	IDF of the lemmatized terms of intermediate concept node B
C_C(IDF)	IDF of the lemmatized terms of destination concept node C
r_1(seen)	Number of papers that contain the influence relation A $\to$ B
r_2(seen)	Number of papers that contain the influence relation B $\to$ C

Table 6.3: List of features used by the link prediction models to classify the candidate link $A \to C$, given an intermediate node B. The top portion of table lists the features derived from the influence graph. The middle lists features extracted from the citation graph. The bottom portion contains features extracted from the collection of papers.

Results

Table 6.2 records the performance of several classifiers evaluated on the test partition,[7] compared against two baselines.

The first baseline randomly creates positive links following the distribution of positive examples from the training partition, while the second baseline assumes that all candidate links are positive (i.e., candidate $A \to C$ is always correct if $A \to B$ and $B \to C$ exist for some B).

From these results we can make several important observations. First, the performance of the random baseline is very low, suggesting that this is indeed a hard task that is exacerbated by the biased label distribution. The precision of the second baseline is similarly low, confirming that the transitive closure assumption is not supported in the influence graph. All of the classifiers examined considerably outperform the baseline, indicating that capturing the structure of both the influence and citation graphs [8] is indeed telling of the likelihood that an influence relation will be discovered in the near future. Interestingly, a support vector machine (SVM) with a linear kernel did not converge on this data, which suggests that while it is possible to learn a model for this link prediction task, the resulting model is more complex than a linear function can capture. Our top-performing model, AdaBoost, is a non-linear classifier which obtained an F1 score of 27% and mean average precision (MAP) of 68%. This is an order of magnitude higher than either baseline. We include precision at 10 and MAP to highlight the practical value of the approach, and emphasize that a subject matter expert reviewing a short list of positive class predictions has a high chance of encountering a true positive[9].

[7] All classifier hyperparameters were tuned on the development partition, and the classification results were averaged over 5 runs using different random seeds.

[8] For further evidence, see the results of our ablation experiment which we have summarized in Table 6.2.

[9] A prediction that is a viable opportunity for further research

Model performance

	F1	Precision (%)	Recall (%)	P@10	MAP
Baseline (random)	2	2	2	-	-
Baseline (all positive)	3.5	1.8	**1**	-	-
Multilayer perceptron	**27**	39.8	20.6	80	53.7
AdaBoost	27	**53.6**	17.8	**90**	**68.1**
Random Forest	23	24.4	**21.6**	50	30.9

Table 6.4: Unranked scores (precision, recall, and F1), ranked precision (P@10), and mean average precision (MAP), of several classifiers' predictions of future influence relations using backtesting at time $t = 2012$. The baseline predicts that every $A \to C$ link will be discovered after time t, if $A \to B$ and $B \to C$ exist before time t.

Table 6.2 shows the results of an ablation experiment in which we measured the drop in performance when each feature was individually removed from our top-performing model (AdaBoost). This experiment indicates that, importantly, both the influence and citation graphs contribute to the overall performance. Removing individual features from either group impacts performance. Several features have a high impact, including C_{between}(avg-IDF), shortest_path_length, which are extracted from the influence graph, and max P($p_{A\to B}$,$p_{B\to C}$) and Jaccard($\mathbf{p}_{A\to B}$,$\mathbf{p}_{B\to C}$), which are extracted from the citation graph. These results demonstrate that the task of scientific discovery requires a multi-faceted approach that analyzes several graphs, including graphs that model the content of publications (our IG), as well as citation graphs.

Ablation experiment

Removed Feature	F1 (%)	Precision (%)	Recall (%)
Full model	26.8	52.8	17.6
− C_A(outdegree)	23.4	67.8	14
− C_A(indegree)	24.6	44	17
− C_C(outdegree)	21.4	24.8	18.6
− C_C(indegree)	20	23.2	17.2
− C_{between}(outdegree)	22	27.2	18.2
− C_{between}(indegree)	23.4	30	19.2
− C_{between}(avg-idf)	21.4	27.2	18
− r_{between}(avg-seen)	23	29	19.2
− shortest_path_count	22.2	29	18.2
− shortest_path_length	20.4	28	16
− max P($p_{A\to B}$,$p_{B\to C}$) (c=100)	22.6	30	18
− min P($p_{A\to B}$,$p_{B\to C}$) (c=100)	22.8	30.2	18.4
− avg P($p_{A\to B}$,$p_{B\to C}$) (c=100)	23.2	30.6	18.6
− max P($p_{A\to B}$,$p_{B\to C}$) (c=300)	23.2	31.4	18.2
− min P($p_{A\to B}$,$p_{B\to C}$) (c=300)	23	29.8	18
− avg P($p_{A\to B}$,$p_{B\to C}$) (c=300)	23.2	32.6	18.2
− Jaccard($\mathbf{p}_{A\to B}$,$\mathbf{p}_{B\to C}$)	22.6	29	18.4
− Inter-citation ratio	23	31.8	18
− C_A(IDF)	24.8	47.8	16.8
− C_B(IDF)	21.6	25.8	19
− C_C(IDF)	22	25.6	19.4
r_1(seen)	22.6	28.4	19
r_2(seen)	22.8	29.8	18.4

Table 6.5: Results of an ablation experiment where one feature was removed at a time from the full AdaBoost model. This experiment was performed on the development partition.

As a final experiment, we rank the discoveries predicted by our model using a scoring function for informativeness that combines classifier confidence and redundancy (i.e., the number of time $A \to C$ was observed with different intermediary nodes B) using a variation of Noisy-OR:

$$Score(A \to C) = (1 - \prod_{B}(1 - prob(A \to C|B))) \tag{6.1}$$

In this version, B loops over all intermediate nodes that support the predicted relation. $prob(A \to C|B)$ denotes the classifier's output probability given a single intermediary node B[10]. Table 6.2 lists the top 10 predictions of our approach under our custom scoring function. 80% of these predictions are correct[11]. We can see that the ranked predictions are indeed informative: they capture fragments of protein signaling pathways, as well as links to biological processes (e.g., apoptosis). A few predicted links such as "TLR → cascade" are not informative, but we largely attribute this to limitations in the machine reader, which failed to capture meaningful content from the destination concept ("cascade").

[10] This ranking function was also used to generate the ranked evaluation in Table 6.2.
[11] They are reported in the influence graph after time $t = 2012$

Selected predictions

Predicted discovery	**Score**	Gold label
antibodies → apoptosis	1	1
apoptosis → ROS	1	1
TGF-beta → apoptosis	1	1
TLR → cascade	1	1
apoptosis → insulin	1	0
apoptosis → enzymes	1	0
antibodies → receptor	1	1
IL-6 → tumor	1	1
mutations → inflammation	0.99	1
macrophages → tumor	0.99	1

Table 6.6: The top 10 influence relations predicted by our model, sorted in descending order of their informativeness score.

6.3 *Bridging distant research communities*

In the previous section, we examined how an influence graph can be combined with a citation graph to predict future scientific discoveries. Next we will explore how a citation graph can be aligned with an influence graph to map the dispersion of scientific findings and estimate the degree of communication taking place between communities. Research draws upon other work and exists in some larger context. In its dissemination, there is an opportunity and hope that research will inform and influence subsequent work. The flow of ideas can be glimpsed in how research cites and is cited by other work. While the potential reach of an idea is a matter of its intrinsic merit and interdisciplinarity, its actual dispersion may be a matter of its framing and the venue in which it is released. As researchers, we give careful thought to the choice of venues we target for publication. Our strategy may be based on our perception of the venue's audience, its repute, or some combination of these factors. Together these elements define a notion of *suitability*[12] which researchers use to rank and select venues for the submission process. By analyzing a large network of citations, we can derive venue-level representations which can be used to measure the exchange of ideas between them. When considered collectively, these venue representations can help to delineate the boundaries of still larger communities. With a reliable means of estimating a venue's awareness of the work published in another venue, we can surface chains of influence relations which may bridge seemingly isolated research communities.

[12] In most cases, it is not only the preference of the authors, but also the assessment of their peers which determines a work's ultimate acceptance and (as a consequence) its immediate audience.

Approach

The influence relations which comprise the graph described in Hahn-Powell et al. (2017) were mined from over 100K open access publications retrieved from PubMed Central[13]. Given that the aim of this work is to generate reliable venue representations and align them to each influence relation, our citation graph must minimally include the venues that published these papers. From Luo et al. (2018) we know that the citation network mirroring this set of publications covers roughly 5.5 million distinct publications. To maximize coverage with freely available resources, I elected to mine all of PubMed, the National Library of Medicine's index of publications from the biomedical and life sciences. The index includes both MEDLINE[14], a collection of metadata for over 25 million publications, and PubMed Central (PMC), a repository over 2 million[15] full-text publications. For consistency and simplicity in this proof-of-concept, only citations for papers with a PubMed identifier (PMID) were considered for inclusion. The vast majority of MEDLINE abstracts did not include usable citation links. Of the 28 million publications mined in this work, only 4 million papers had citations to and from papers possessing a PMID; however, these 4 million papers yielded 124.9 million usable citation links.

[13] `https://www.ncbi.nlm.nih.gov/pmc/`

[14] See `https://www.nlm.nih.gov/bsd/medline.html`. Over 80% of these entries in MEDLINE contain abstracts.

[15] The official reported count as of May 2018: `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

Each PMID was mapped to its venue's name[16] In total, this set of links involves **19,373** distinct venues. For each venue, I created a 38,746 dimensional vector of counts. Half of the dimensions counted incoming venue-level citations, while the other half was reserved for outgoing venue citations. A toy citation network is depicted in Figure 6.3 to illustrate how the outgoing citation counts were amassed.

[16] For each link, I also retained the year of publication with the intention of generating venue representations using the links leading up and through each year in the dataset in order to explore the temporal aspect of shifting community boundaries.

Because the resultant vectors are extremely sparse, I applied truncated Singular Value Decomposition [17] (SVD) to generate a low-rank

[17] Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN 0521865719 9780521865715

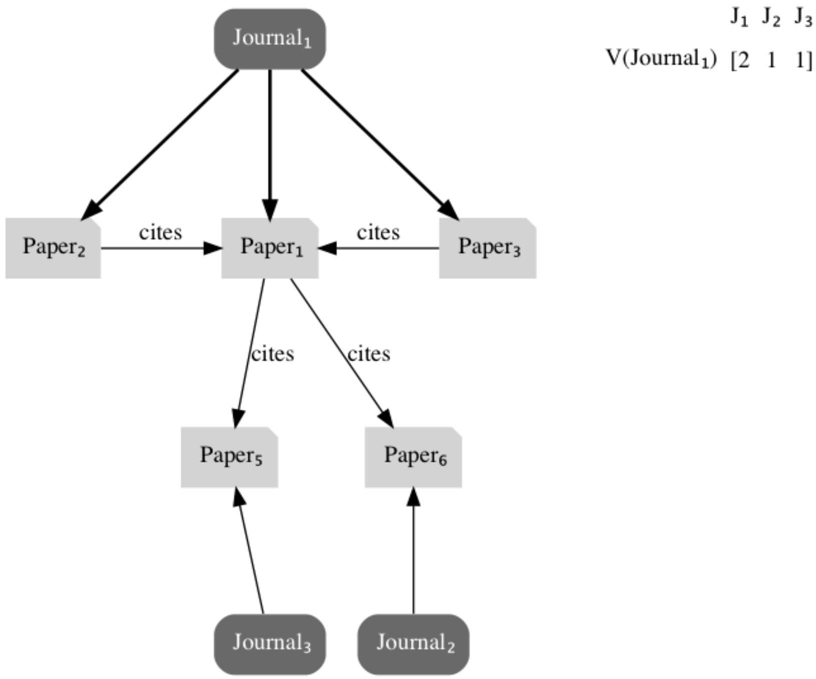


Figure 6.2: An example citation network illustrating how the outgoing citations of a venue (represented here as Journal_1) were counted. Each journal publishes papers. These papers cite other papers, each of which was published in some venue.

approximation of k dimensions. In this case, a value of 100 for k was sufficient to account for 99% of the observed variance in the original 38,746-dimensional sparse matrix M. The rows of the product matrix C (see Figure 6.3) are low-rank dense representations corresponding to each of the 19,373 venues. Aside from gaining a compressed representation which retains the relative hyperspace positions of the original sparse vectors, the dense representations produced by SVD have an additional advantage in that each dimension corresponds to some latent feature[18]. Moreover, the dimensions are produced in decreasing order of explained variance.

We can assess the quality of these representations in several ways. As a sanity check, we can first cluster these elements and examine them to determine whether or not the groupings possess the desired semantics. Figures 6.3 and 6.3 show selected portions of a two-dimensional GMap (Gansner et al., 2010) clustering of these venues. In this case, the map is parameterized to limit cluster membership to groups of venues meeting an aggressive cosine similarity threshold of 0.97. The clusters have good cohesion. In Figure 6.3, we see a clear cluster of venues related to urology research, even though the method includes no explicit content analysis of the publications found in each of these venues.

[18] For example, dimension 1 of each dense representation encodes a different value for the same latent feature.

$$C_k = U \Sigma V^T$$

Figure 6.3: Equation for Singular value Decomposition. Given a matrix M, a low-rank least squares approximation C is defined as the product of three matrices: U is an $m \times k$ matrix, V is an $n \times k$ matrix, and Σ is a square $k \times k$ diagonal matrix where the singular values along the diagonal reflect the importance (e.g., explained variance) of each latent dimension (column) in U and V. In this formulation, U and V are orthonormal matrices. k corresponds to the rank (i.e., number of rows) of the least squares approximation matrix C.

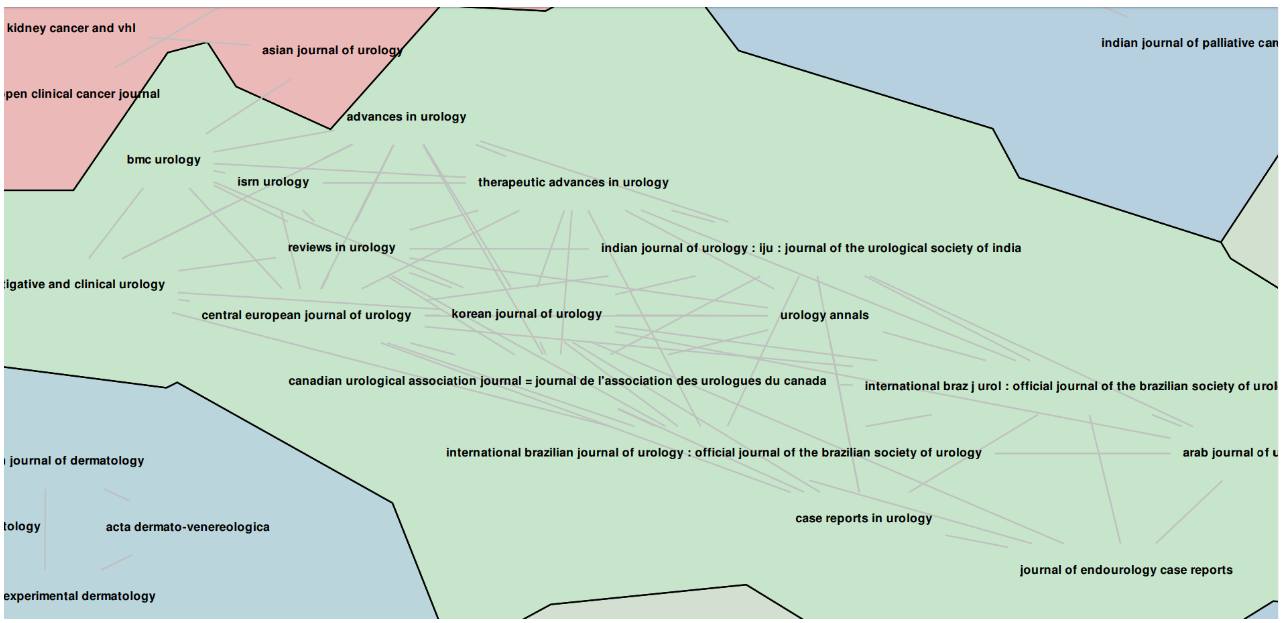


Figure 6.4: A portion of a GMap (Gansner et al., 2010) visualization depicting clusters of venues related to urology and dermatology.

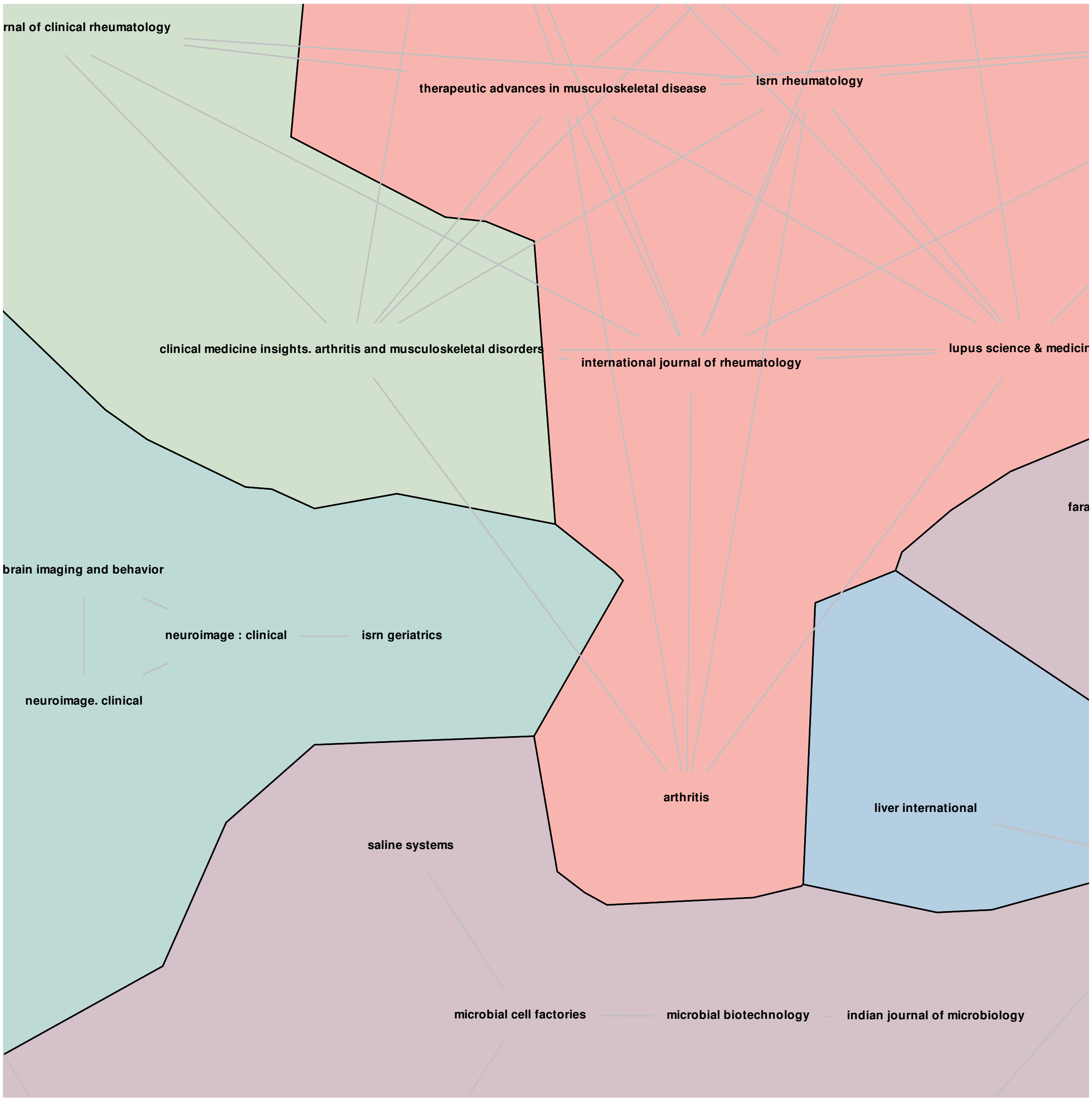


Figure 6.5: Another portion of the same GMap visualization, this time centered around clusters related to rheumatology, neuroimaging, and microbiology. Grey lines indicate closely related venues, while clusters (or territories) demarcate highly similar groups of venues.

Undiscovered public knowledge

The true test in determining the value of these dense representations is to apply them to some downstream task. Can venue similarity be used to uncover "undiscovered public knowledge" (Swanson, 1986b)? In the 1980s, Don Swanson, the pioneer of literature-based discovery, demonstrated that simply by reading the titles of papers originating from different research communities, one could propose novel treatments to diseases that had been overlooked by researchers. The classic example, depicted in Figure 6.3, is Swanson's observation (Swanson, 1986a) that dietary fish oil, which had been reported to reduce blood viscosity, could be used to treat Reynaud's disease based on independent observations in the literature related to elevated blood viscosity among the afflicted. Not only had the promising link been missed by researchers, the hypothesized treatment proved valid.

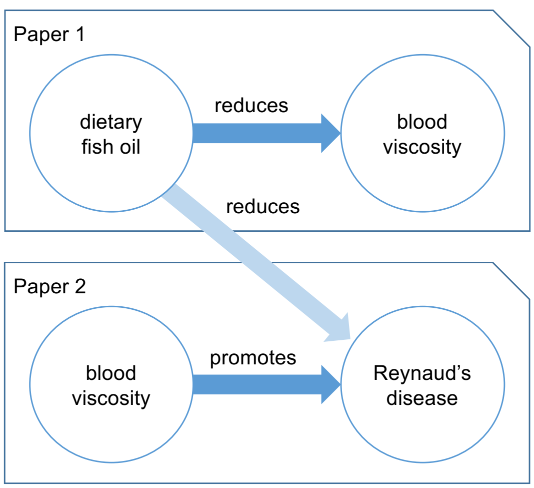


Figure 6.6: Swanson's (1986a) pioneering example of literature-based discovery in which he hypothesized that dietary fish oil could be used to treat Reynaud's disease based on their mutual relation to blood viscosity. The key observations serving as motivation for the proposed treatment were found in the titles of seemingly unrelated papers.

Swanson's earliest discoveries were made through manual analysis; however, the body of scientific literature has grown by an order of magnitude since that time[19]. By using community distance to rank *paths* of influence relations mined from text, we can unearth a potential trove of novel hypotheses synthesized from these disparate findings. Because each influence relation is supported by evidence from one or more papers, we can measure community cohesion both within and across relations by using a distance metric that compares the venue vectors corresponding to each paper. Figure 6.3 summarizes the approach. Here I use cosine similarity as the distance metric to compute an average of pairwise similarity scores for all venues

[19] There were roughly 9 million publications between 1800-1989. Since 1990, the total has grown by an additional 19 million publications.

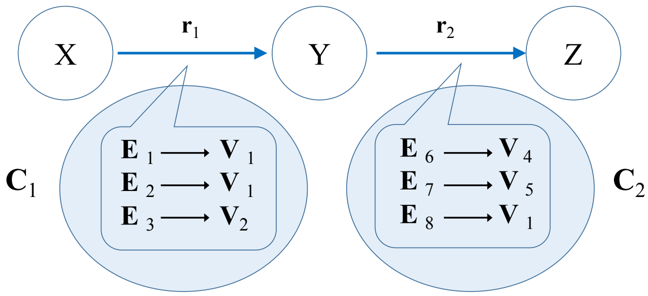


Figure 6.7: A summary of the alignment of venues to influence relations. r_1 and r_2 are relations in the influence graph linking concepts X, Y, and Z. Both r_1 and r_2 are supported by evidence (E_n) mined from the literature. Each piece of evidence originates from a publication disseminated in some venue (V_n). Together, these venues form a community (C_n) or set of communities. We can measure the community distance of r_1 and r_2 by comparing their associated communities using venue-level representations. The current work uses an average of pairwise distances for the involved venues (e.g., $Ave(sim(V_1, V_4), sim(V_1, V_5), \ldots, sim(V_2, V_1))$). The higher the score, the greater the community cohesion of the involved venues.

in an influence relation path[20] of length 2. I use a filtering criteria in the style described in Luo et al. (2018), though only to highlight novel concepts through a minimum IDF score without imposing any redundancy criteria. This is to ensure that paths involving distant communities also contain concepts that are somewhat uncommon overall in the graph. What follows are selected results with low community similarity for a variety of indirect queries. This variety is intended to test the generalizability of the technique.

[20] While chains of length 2 were used here, the proposed method generalizes to influence paths of any length.

Parkinson's disease (1 of 2)

$\mathbf{r_1}$: *acquired mitochondrial dysfunction* INCREASES *oxidative stress*
$\mathbf{r_2}$: *oxidative stress* INCREASES *Parkinson models*

Venues	*acta psychiatrica scandinavica*	*acta neuropathologica communications*
Similarity	**0.0058**	

Table 6.7: Influence pathway linking acquired mitochondrial dysfunction to Parkinson's through their mutual relationship to oxidative stress. The evidence for this path was published in 2014. In 2015, a paper linking *acquired mitochondrial dysfunction* with *Parkinson's* was published: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4479806`

Parkinson's disease (2 of 2)

$\mathbf{r_1}$: *Akt phosphorylation* INCREASES *angiogenesis*
$\mathbf{r_2}$: *angiogenesis* DECREASES *Parkinson's disease*

Venues	*journal of ophthalmology*	*journal of clinical psychopharmacology*
Similarity	**0.0067**	

Table 6.8: Influence pathway where the first relation describes how common viruses can interfere with AKT phosphorylation, which plays a role in angiogenesis (the development of new blood vessels). Separately, a link between angiogenesis and Parkinson's is suggested, though subsequent research has found that excessive angiogenesis (cf. impaired) may in fact be a cause of Parkinson's. A potential link between angiogenesis and Parkinson's is an opportunity for further research.

Pneumonia (1 of 2)

r_1: *rifabutin* DECREASES *clarithromycin*
r_2: *clarithromycin* DECREASES *the severity of the pneumonia*

Venues	*infection*	*mediators of inflammation*
Similarity	**0.0378**	

Table 6.9: Influence pathway describing a drug interaction where one antibiotic (rifabutin) used to treat tuberculosis can interfere with the efficacy of another (clarithromycin), which is used to treat pneumonia.

Pneumonia (2 of 2)

r_1: *C. auris/M. pachydermatis* INCREASES *fungemia*
r_2: *fungemia* INCREASES *pneumonia*

Venues	*emerging infectious diseases*	*clinics in orthopedic surgery*
Similarity	**0.026**	

Table 6.10: The presence of fungal spores in the blood may increase risk of developing pneumonia.

Autism (1 of 3)

r_1: *Beta-catenin translocation* INCREASES *Wnt*
r_2: *Wnt* INCREASES *autism*

Venues	*journal of hepatology*	*journal of neurodevelopmental disorders*
Similarity	**0.0316**	

Table 6.11: Beta-catenin translocation may affect Autism by means of the Wnt pathway.

Autism (2 of 3)

r_1: *HBV X protein* INCREASES *Wnt*
r_2: *Wnt* INCREASES *autism*

Venues	*journal of hepatology*	*journal of neurodevelopmental disorders*
Similarity	**0.0316**	

Table 6.12: A protein produced by the Hepatitis virus may have a connection to Autism through its effect on the Wnt pathway.

Autism (3 of 3)

r_1: *adequate sunlight* INCREASES *vitamin D*
r_2: *vitamin D* DECREASES *autism symptom severity*

Venues	[*environmental health perspectives*, *bmc psychiatry*, *scientific reports*]	*dermato-endocrinology*
Similarity	**0.234**	

Table 6.13: Influence pathway highlighting growing evidence of a link between vitamin D and Autism. Reportedly, vitamin D supplementation can reduce the severity of Autism symptoms.

Meningitis

r_1: *resveratrol from bilberries* and *Nettle leaf extracts* DECREASES *NF-kappaB*
r_2: *NF-kappaB* INCREASES *bacterial meningitis*

Venues	*journal of ethnopharmacology*	*frontiers in microbiology*
Similarity	**0.00275**	

Table 6.14: Through their effect on NF-kappaB, bilberries and nettle leaf extract may have some potential to treat or prevent bacterial meningitis.

Neuroinflammation

r_1: *A. terreus* INCREASES *lovastatin*
r_2: *lovastatin* DECREASES *neuroinflammation*

Venues	*alimentary pharmacology & therapeutics*	*malaria journal*
Similarity	**0.031**	

Table 6.15: Certain fungi produce lovastatin, which can decrease neuroinflammation. This has potential implications in the treatment protocols for a variety of diseases such as malaria and Alzheimer's.

Cannabidiol

r_1: *CBD* DECREASES *NF-kappaB*
r_2: *NF-kappaB* INCREASES *renal fibrosis*

Venues	*journal of neuroinflammation*	*urology annals*
Similarity	**0.058**	

Table 6.16: CBD, one of the non-psychoactive compounds found in cannabis, may serve as a preventative supplement for kidney disorders through its effect on NF-kappaB.

THC

r_1: *THC* (**but NOT Cannabidiol**) DECREASES *IOP*
r_2: *IOP* INCREASES *ocular hypertension*

Venues	*dialogues in clinical neuroscience*	*journal of current glaucoma practice*
Similarity	**0.0257**	

Table 6.17: THC, the primary pychoactive compound in cannabis, may have a therapeutic effect on treating ocular hypertension by temporarily alleviating inter ocular pressure (IOP), though the same effect does not hold true for the nonpyschoactive compound Cannabidiol (CBD).

Diabetes (1 of 3)

r_1: *sleep fragmentation* DECREASES *insulin sensitivity*
r_2: *reduced insulin sensitivity* INCREASES *diabetes*

Venues	*european journal of endocrinology*	[*journal of animal science and biotechnology*, *trials*]
Similarity	**0.179**	

Table 6.18: Sleep fragmentation can decrease insulin sensitivity and thus increase the risk of developing diabetes.

Diabetes (2 of 3)

r_1: *niacin* DECREASES *insulin resistance*
r_2: *insuline resistance* INCREASES *diabetes*

Venues	[*bmc public health*, *clinical medicine insights: cardiology*]	*plos one*
Similarity	**0.208**	

Table 6.19: Niacin, which is sometimes recommended to help manage cholesterol, can increase insulin resistance and therein raises one's risk of developing diabetes.

Diabetes (3 of 3)

r_1: *clozapine* INCREASES *insulin*
r_2: *insulin* INCREASES *risk of cardiovascular disease and diabetes*

Venues	[*neuropsychopharmacology*, *plos genetics*]	[*international journal of endocrinology and metabolism*, *bmc women's health*]
Similarity		

Table 6.20: A drug used to treat schizophrenia, clozapine, has an effect on insulin signaling, which increases the risk of developing both cardiovascular disease and diabetes.

6.4 Conclusions and limitations

Future discoveries In this chapter I have described how a conceptual influence graph can be used to learn to predict likely discoveries in the near future. While our initial results are modest, the influence graph used is relatively small and was intended to only address issues surrounding children's health. Results would likely improve by increasing the size and scope of the influence graph, as well as exploring alternative models of research communities.

Research community interaction By pairing a large citation graph with an influence graph, we can surface opportunities for collaboration and discovery by identifying disconnected research communities. I presented a variety of influence relation pathways that cross the boundaries of distant pairs of communities [21]. Until the results are validated and verified by members of the research communities in question, the value of this approach remains largely unproven. Moreover, while the general approach appears promising, its usefulness may be limited both by the quality of community representations and that of the machine reading system used to identify the influence relations.

[21] Communities which have had little observed interaction and where it is thus expected that members of either community are unlikely to be aware of related discoveries known to the other.

7 Conclusions

7.1 Summary

In this work, I have argued that large-scale machine reading is essential to accelerating scientific progress, but existing approaches are hampered by a limited ability to contextualize and synthesize individual facts mined from text.

To help address the challenge of "undiscovered public knowledge" (Swanson, 1986a; Swanson and Smalheiser, 1996; Bekhuis, 2006), I have demonstrated how text mined fragments can be assembled and causally ordered in a domain specific setting to extend our understanding of mechanistic biology. Deep, targeted approaches to machine reading (Valenzuela-Escárcega et al., 2017) can be generalized and extended to cover a confluence of domains where individual statements can be pieced together automatically to form a conceptual influence graph of linked assertions each of which is substantiated by structured textual spans from one or more publications. I presented evidence that such graphs can inform and greatly accelerate model development. They can also help to identify and eliminate bias.

By aligning this knowledge graph with a large network of citations, we can begin to identify the boundaries of research communities and derive semantic representations of these communities based on their interaction. These representations help power a form of community-aware machine reading capable of uncovering chains of findings that bridge the work of distant communities using research originating from both sides of the division. Grounded in prior insights, the synthesis of neighboring findings makes for a compelling form of hypothesis generation. This work, however, is far from over. A true solution will require interdisciplinary thinking and a concerted effort of AI researchers, linguists, and subject matter experts.

7.2 Future work

I limit discussion of future work to the following three primary topics.

Community-aware machine reading Chapter 6 demonstrated a proof-of-concept for identifying research communities based on their interaction in a citation graph. A simple extension of this work would be to apply a local context window model such as skip-gram (Mikolov et al., 2013b) or the global log-bilinear regression model of Pennington et al. (2014) to generate alternative semantic representations of each venue and compare their relative similarity to that of the set of dense representations developed in this work. Beyond improvements to the quality of these representations, though, are new applications that make use of a measure of their interaction. Can, for example, the result of a query to the influence graph be tailored to a researcher's background (e.g., the papers they have published)? Can we learn to predict the contributions of co-authors using this same technique? What can such an approach tell us about the composition of organizations? Similarly, it may be revealing to understand the number of communities and the distribution of their involvement in a particular research topic. New applications aside, the example hypotheses presented in this work need to somehow be verified or (minimally) assessed for their potential value. One option may be to contact the authors of the foundational papers supporting the relations that compose the hypothesis, as they are likely best-suited to evaluating these hypotheses.

Assembly The causal precedence work described in Chapter 3 can most readily be improved by increasing the size of corpus. At some potential cost to sample diversity, this process can be expedited by running the current best k models in an active learning loop. Ideally, this effort could be crowd-sourced to a community of domain experts. One exciting possibility for improving the coverage of influence graphs described in Chapter 4, is to construct a multilingual alignment of various monolingual knowledge graphs. Such an extension is already underway with Portuguese and Spanish using content repositories like SciELO[1]. We hope that insights arising from this kind of assembly will help to foster meaningful collaboration across research communities.

[1] http://www.scielo.org

Guided reading Improving the precision of machine readers, especially mixed or domain-independent approaches, can be achieved in several complementary ways. Ontology acquisition is a promising, albeit challenging, avenue of research for guided reading. Learning to selectively apply existing ontologies (Kalfoglou and Schorlemmer, 2003; Tsai and Roth, 2016) or infer a hierarchical structure (Steeg and Galstyan, 2014) from the content of papers will help to eliminate noise and enable sophisticated reasoning over reader output.

The quality of concepts identified in mixed or domain-independent systems has a major downstream effect on the capabilities of entity linking and assembly. A much easier means of enhancing the quality of reading in the short-term is in assessing the reliability of claims in the literature by detecting telling features like P-values, sample sizes, and by classifying the type of experimental design. Additionally, detecting whether or not a paper is accompanied by a dataset and understanding the variables and measurements captured in that data may help to estimate a study's reproducibility. The same applies to interpreting tables and figures (Clark and Divvala, 2015, 2016).

A Reach

Table A.1

A.1 Assembly annotation schema

The following tables provide a description of the annotation schema used by subject matter experts to annotate state and precedence information. The expert feedback (gold data) was used to inform the development of the assembly component.

INPUT	One or more entities or Events serving as input. In the case of an Event, use an `EVENT ID`. When the output is an Event, do not list any status changes). Multiple inputs should be delimited with `,` (ex. `A,B`).
	Modifications on entities (status changes) should be specified with `.` (ex. `ENTITY.mod1.mod2`) following the entity.
	modifications:
	`u = ubiquitination`
	`p = phosphorylation`
	`...`

OUTPUT	One or more entities or Events produced as output. In the case of an Event, use an `EVENT ID`. When the output is an Event state information is not appended to the ID. Multiple inputs should be delimited with , (ex. `A,B`). Modifications on entities (status changes) should be specified with . (ex. `ENTITY.mod1.mod2`) following the entity.
CONTROLLER	The controller or "cause" of some event. Note that this could be an event (ex. E2). The "sign" of a controller is positive, unless stated otherwise.
PRECEDED BY	The list of `EVENT ID`s that implicitly precede this event.
EVENT ID	Events should be labeled `E1` ... `EN`, where N is the total number of distinct events; The event selected for E1 need not be the first event detected.

CONFIDENCE	Whenever an entry is ambiguous, please label it using one of three following confidence levels:
	`high`
	`medium`
	`low`
	If no level is specified, it is assumed to be "high".

Table A.1: A description of how each column's data should be encoded for assembly annotations.

INPUT	OUTPUT	CONTROLLER	PRECEDED BY	EVENT ID	TEXT
A	A.a	D		E1	D activates A
B	B.p	C		E2	A inhibits the phosphorylation of B by C.
E2	E2	A.a.neg		E3	A inhibits the phosphorylation of B by C.
B	B.p	E		E4	B is also phosphorylated by E.
B.p	B.u.p	C	E4	E5	Then, B is ubiquitinated by C.
X	X.u	Y	E5	E6	Then, X is ubiquitinated by Y.
X.u,B.p	{X.u,B.p}			E7	Ubiquitinated X binds to phoshorylated B
B	B.a	X[s]		E8	The X mutants (s, t, and v) activate B
B	B.a	X[t]		E9	The X mutants (s, t, and v) activate B
B	B.a	X[u]		E10	The X mutants (s, t, and v) activate B
B	B.a	X[t].p		E11	Phosphorylated Xt activates B.
B	B.p@z	E		E12	B is also phosphorylated by E at z.

Table A.2: An annotation example for "D activates A. A inhibits the phosphorylation of B by C. B is also phosphorylated by E. Then, B is ubiquitinated by C. Then, X is ubiquitinated by Y. Ubiquitinated X binds to phoshorylated B." In addition to the fields listed here, annotators also assigned confidence scores on a 3-point likert scale (high, medium, low) to record cases of textual ambiguity. Confidence was assumed to be high if the field was left unmarked.

A.2 Example assembly output

Running Reach with `outputType = "assembly"` produces a tab-delimited file with a superset of the columns listed in Table A.1.[1]

Each row corresponds to a distinct (i.e., deduplicated) reaction.

Table A.1 provides sample assembly output for the following two sentences taken from PMC1242134[2]:

> Akt interacts with Raf and phosphorylates it at Ser259 . Furthermore , phosphorylation of Raf by Akt inhibits activation of the Raf-MEK-ERK signaling pathway and has been shown to alter the cellular response in a human breast cancer cell line from cell cycle arrest to proliferation .

[1] Additional columns related to context are included in the output, but omitted from Table A.2, since they were unpopulated for this example.

[2] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1242134/`

INPUT	OUTPUT	CONTROLLER	EVENT ID	EVENT LABEL	NEGATED	INDIRECT?	CONTEXT (SPECIES)	CONTEXT (TISSUE TYPE)	SEEN	EVIDENCE	SEEN IN
Akt::uniprot:P31749, Raf::uniprot:P04049	{Akt::uniprot:P31749, Raf::uniprot:P04049}	NONE	E11	Binding	false	true	taxonomy:9606	tissuelist:TS-0591	1	Akt interacts with Raf and phosphorylates it at Ser259 .	PMC1242134
Raf::uniprot:P04049	Raf::uniprot:P04049.p@Ser259	NONE	E3	Phosphorylation	false	true	taxonomy:9606	tissuelist:TS-0591	1	Akt interacts with Raf and phosphorylates it at Ser259 .	PMC1242134
Raf::uniprot:P04049	Raf::uniprot:P04049.p	NONE	E8	Phosphorylation	false	false			1	Furthermore , phosphorylation of Raf by Akt inhibits activation of the Raf-MEK-ERK signaling pathway and has been shown to alter the cellular response in a human breast cancer cell line from cell cycle arrest to proliferation [XREF_BIBR] .	PMC1242134
E3	Raf::uniprot:P04049.p@Ser259	Akt::uniprot:P31749	E4	Regulation	false	false	taxonomy:9606	tissuelist:TS-0591	1	Akt interacts with Raf and phosphorylates it at Ser259 .	PMC1242134
E8	Raf::uniprot:P04049.p	Akt::uniprot:P31749	E9	Regulation	false	false	taxonomy:9606	tissuelist:TS-0591	1	Furthermore , phosphorylation of Raf by Akt inhibits activation of the Raf-MEK-ERK signaling pathway and has been shown to alter the cellular response in a human breast cancer cell line from cell cycle arrest to proliferation [XREF_BIBR] .	PMC1242134

Table A.3: Example assembly output for two sentences taken from PMC1242134: *Akt interacts with Raf and phosphorylates it at Ser259 . Furthermore , phosphorylation of Raf by Akt inhibits activation of the Raf-MEK-ERK signaling pathway and has been shown to alter the cellular response in a human breast cancer cell line from cell cycle arrest to proliferation .*

References

(2012). TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. *Expert Systems with Applications*, **39**(3), pp. 2927 – 2938. ISSN 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2011.08.154.

Afolabi, I., G. Musa, C. Ayo, and A. Sofoluwe (2008). Knowledge discovery in online repositories: a text mining approach. *European Journal of Scientific Research*, **22**(2), pp. 241–250. ISSN 1450-216X.

Allen, J. F., M. Swift, and W. de Beaumont (2008). Deep Semantic Analysis of Text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pp. 343–354. Association for Computational Linguistics, Stroudsburg, PA, USA.

Appelt, D. E., J. R. Hobbs, J. Bear, D. Israel, and M. Tyson (1993). FASTUS: A finite-state processor for information extraction from real-world text. In *IJCAI*, volume 93, pp. 1172–1178.

Appelt, D. E. and B. Onyshkevych (1998). The Common Pattern Specification Language. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, pp. 23–30. DOI: 10.3115/1119089.1119095.

Babur, Ö., M. Gönen, B. A. Aksoy, N. Schultz, G. Ciriello, C. Sander, and E. Demir (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, **16**(1), p. 45. ISSN 1465-6906. DOI: 10.1186/s13059-015-0612-6.

Banko, M., M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007). Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 2670–2676. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, **3**(1), p. 2. ISSN 1742-5581. DOI: 10.1186/1742-5581-3-2.

Bell, D., G. Hahn-Powell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). An investigation of coreference phenomena in the biomedical domain. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC.

Berant, J., A. Chou, R. Frostig, and P. Liang (2013). Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544. Association for Computational Linguistics, Seattle, Washington, USA.

Berant, J., V. Srikumar, P.-C. Chen, A. Vander Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning (2014). Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499–1510. Association for Computational Linguistics. DOI: 10.3115/v1/D14-1159.

Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio (2010). Theano: a CPU and GPU Math Expression Compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation.

Bethard, S., L. Derczynski, G. Savova, J. Pustejovsky, and M. Verhagen (2015). SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 806–814. Association for Computational Linguistics. DOI: 10.18653/v1/S15-2136.

Bethard, S. and J. H. Martin (2008). Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations. In *Proceedings of ACL-08: HLT, Short Papers*, pp. 177–180. Association for Computational Linguistics.

Bethard, S., G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen (2016). SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1052–1062. Association for Computational Linguistics. DOI: 10.18653/v1/S16-1165.

Björne, J., J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski (2009). Extracting Complex Biological Events with Rich Graph-Based Feature Sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 10–18. Association for Computational Linguistics.

Björne, J. and T. Salakoski (2013). TEES 2.1: Automated Annotation Scheme Learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 16–25. Association for Computational Linguistics.

Blaschke, C., M. A. Andrade, C. Ouzounis, and A. Valencia (1999). Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 60–67. AAAI Press. ISBN 1-57735-083-9.

Bui, Q.-C. and P. M. Sloot (2012). A robust approach to extract biomedical events from literature. *Bioinformatics*, **28**(20), pp. 2654–2661. DOI: 10.1093/bioinformatics/bts487.

Califf, M. E. and R. J. Mooney (1997). Relational Learning of Pattern-Match Rules for Information Extraction. In *CoNLL97: Computational Natural Language Learning*.

Califf, M. E. and R. J. Mooney (2003). Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*, **4**, pp. 177–210. DOI: 10.1162/153244304322972685.

Cameron, D. H. (2015). A Context-driven Subgraph Model for Literature-based Discovery. *SIGWEB Newsl.*, (Winter), pp. 2:1–2:2. ISSN 1931-1745. DOI: 10.1145/2719943.2719945.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell (2010). Toward an Architecture for Never-Ending Language Learning.

Cerami, E. G., B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, pp. D685–D690. DOI: 10.1093/nar/gkq1039.

Chambers, N., T. Cassidy, B. McDowell, and S. Bethard (2014). Dense Event Ordering with a Multi-Pass Architecture. *Transactions of the Association of Computational Linguistics*, **2**, pp. 273–284.

Chang, A. X. and C. D. Manning (2014). TokensRegex: Defining cascaded regular expressions over tokens. Technical Report CSTR 2014-02, Computer Science, Stanford.

Chang, C.-C. and C.-J. Lin (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, pp. 27:1–27:27. ISSN 2157-6904. DOI: 10.1145/1961189.1961199. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chiticariu, L., R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan (2010). SystemT: An Algebraic Approach to Declarative Information Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 128–137. Association for Computational Linguistics.

Chiticariu, L., Y. Li, and F. R. Reiss (2013). Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 827–832. Association for Computational Linguistics.

Chollet, F. (2015). keras. `https://github.com/fchollet/keras`.

Clark, C. and S. Divvala (2015). Looking Beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers.

Clark, C. and S. Divvala (2016). PDFFigures 2.0: Mining Figures from Research Papers. pp. 143–152. DOI: 10.1145/2910896.2910904.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), pp. 37–46. DOI: 10.1177/001316446002000104.

Cohen, K. B. and D. Demner-Fushman (2014). *Biomedical Natural Language Processing*, volume 11. John Benjamins Publishing Company. DOI: 10.1075/nlp.11.

Cohen, P. R. (2015). DARPA's Big Mechanism program. *Physical biology*, **12**(4), p. 045008.

Craven, M. and J. Kumlien (1999). Constructing Biological Knowledge Bases by Extracting Information from Text Sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 77–86. AAAI Press. ISBN 1-57735-083-9.

Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan (2002). GATE: an Architecture for Development of Robust HLT applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Dang, T. N., P. Murray, and A. G. Forbes (2015). PathwayMatrix: visualizing binary relationships between proteins in biological pathways. *BMC Proceedings*, **9**(6), p. S3. ISSN 1753-6561. DOI: 10.1186/1753-6561-9-S6-S3.

de Marneffe, M.-C. and C. D. Manning (2008). The Stanford Typed Dependencies Representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Coling 2008 Organizing Committee.

Demir, E., M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'eustachio, C. Schaefer, J. Luciano, et al. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, **28**(9), pp. 935–942. DOI: 10.1038/nbt.1666.

Derczynski, L. and R. Gaizauskas (2013). Empirical Validation of Reichenbach's Tense Framework. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pp. 71–82. Association for Computational Linguistics.

Drozdzynski, W., H.-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu (2004). Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. *Könstliche Intelligenz*, **18**, pp. 17–.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2004). Web-scale Information Extraction in Knowitall: (Preliminary Results). In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pp. 100–110. ISBN 1-58113-844-X. DOI: 10.1145/988672.988687.

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.*, **9**, pp. 1871–1874. ISSN 1532-4435.

Gansner, E. R., Y. Hu, and S. G. Kobourov (2010). GMap: Visualizing graphs and clusters as maps. PacificVis, pp. 201–208. IEEE. DOI: 10.1109/PACIFICVIS.2010.5429590.

Gould, S. J. (1993). Cordelia's dilemma. *Natural History*, **102**(2), pp. 10–15.

Gunning, D., V. K. Chaudhri, P. Clark, K. Barker, S. Y. Chaw, M. Greaves, B. N. Grosof, A. Leung, D. D. McDonald, S. Mishra, J. Pacheco, B. W. Porter, A. Spaulding, D. Tecuci, and J. Tien (2010). Project Halo Update - Progress Toward Digital Aristotle. *AI Magazine*, **31**(3), pp. 33–58. DOI: 10.1609/aimag.v31i3.2302.

Hahn-Powell, G., D. Bell, M. A. Valenzuela-Escárcega, and M. Surdeanu (2016). This before That: Causal Precedence in the Biomedical Domain. In *Proceedings of the 2016 Workshop on Biomedical Natural Language Processing*, pp. 146–155. Association for Computational Linguistics. DOI: 10.18653/v1/W16-2920.

Hahn-Powell, G., M. A. Valenzuela-Escárcega, and M. Surdeanu (2017). Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Software Demonstrations*, pp. 103–108. ACL. DOI: 10.18653/v1/P17-4018.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Hearst, M. A. (1999). Untangling Text Data Mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

Hirschman, L., A. Yeh, C. Blaschke, and A. Valencia (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, **6**(1), p. S1. ISSN 1471-2105. DOI: 10.1186/1471-2105-6-S1-S1.

Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Huang, M., X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li (2004a). Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, **20**(18), pp. 3604–3612. DOI: 10.1093/bioinformatics/bth451.

Huang, M., X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li (2004b). Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, **20**(18), pp. 3604–3612. DOI: 10.1093/bioinformatics/bth451.

Huffman, S. B. (1996). Learning information extraction patterns from examples. In Wermter, S., E. Riloff, and G. Scheler (eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 246–260. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-49738-7. DOI: 10.1007/3-540-60925-3_51.

Hunter, L., Z. Lu, J. Firby, W. A. Baumgartner, H. L. Johnson, P. V. Ogren, and K. B. Cohen (2008). OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC bioinformatics*, **9**(1), p. 78. ISSN 1471-2105. DOI: 10.1186/1471-2105-9-78.

Jurafsky, D. and J. H. Martin (2018). *Speech and Language Processing*. Prentice-Hall, Upper Saddle River, NJ, USA, 3 edition.

Kalfoglou, Y. and M. Schorlemmer (2003). Ontology Mapping: The State of the Art. volume 18, pp. 1–31. Cambridge University Press. ISSN 0269-8889. DOI: 10.1017/S0269888903000651.

Kilicoglu, H. and S. Bergler (2009). Syntactic Dependency Based Heuristics for Biological Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 119–127. Association for Computational Linguistics.

Kim, J.-D., J.-j. Kim, X. Han, and D. Rebholz-Schuhmann (2015). Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task. *BMC Bioinformatics*, **16**(10), p. S3. ISSN 1471-2105. DOI: 10.1186/1471-2105-16-S10-S3.

Kim, J.-D., N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa (2012). The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics*, **13**(11), p. S1. ISSN 1471-2105. DOI: 10.1186/1471-2105-13-S11-S1.

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 BioNLP 2009 Workshop Companion Volume for Shared Task*, pp. 1–9. Association for Computational Linguistics.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, pp. i180–i182. DOI: 10.1093/bioinformatics/btg1023.

Kim, J.-D., T. Ohta, and J. Tsujii (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, **9**(1), p. 10. ISSN 1471-2105. DOI: 10.1186/1471-2105-9-10.

Kim, J.-D., S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii (2011). Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 1–6. Association for Computational Linguistics.

Klügl, P., M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, **22**(1), pp. 1–40. DOI: 10.1017/S1351324914000114.

Lander, A. D. (2010). The edges of understanding. *BMC Biology*, **8**(1), p. 40. ISSN 1741-7007. DOI: 10.1186/1741-7007-8-40.

Landis, J. R. and G. G. Koch (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**(1), pp. 159–174. ISSN 0006341X, 15410420. DOI: 10.2307/2529310.

Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, **39**(4). DOI: 10.1162/COLI_a_00152.

Leskovec, J., D. Huttenlocher, and J. Kleinberg (2010). Predicting Positive and Negative Links in Online Social Networks. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 641–650. ACM, New York, NY, USA. ISBN 978-1-60558-799-8. DOI: 10.1145/1772690.1772756.

Levy, R. and G. Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Liben-Nowell, D. and J. Kleinberg (2007). The Link-prediction Problem for Social Networks. *J. Am. Soc. Inf. Sci. Technol.*, **58**(7), pp. 1019–1031. ISSN 1532-2882. DOI: 10.1002/asi.v58:7.

Luo, F., M. A. Valenzuela-Escárcega, G. Hahn-Powell, and M. Surdeanu (2018). Scientific Discovery as Link Prediction in Influence and Citation Graphs. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pp. 1–6. Association for Computational Linguistics. DOI: 10.18653/v1/W18-1701.

Magrane, M. and U. Consortium (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, p. bar009. DOI: 10.1093/database/bar009.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. Association for Computational Linguistics. DOI: 10.3115/v1/P14-5010.

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press. ISBN 0521865719 9780521865715.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, **19**(2).

McClosky, D., S. Riedel, M. Surdeanu, A. McCallum, and C. D. Manning (2012). Combining joint models for biomedical event extraction. *BMC Bioinformatics*, **13**(11), p. S9. ISSN 1471-2105. DOI: 10.1186/1471-2105-13-S11-S9.

Mihăilă, C., T. Ohta, S. Pyysalo, and S. Ananiadou (2013). BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, **14**(1), pp. 1–18. ISSN 1471-2105. DOI: 10.1186/1471-2105-14-2.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.) *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

Mintz, M., S. Bills, R. Snow, and D. Jurafsky (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011. Association for Computational Linguistics.

Mirza, P. (2014). Extracting Temporal and Causal Relations between Events. In *Proceedings of the ACL 2014 Student Research Workshop*, pp. 10–17. Association for Computational Linguistics. DOI: 10.3115/v1/P14-3002.

Mirza, P. (2016). *Extracting Temporal and Causal Relations between Events*. Ph.D. thesis, University of Trento.

Mirza, P. and S. Tonelli (2016). CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 64–75. The COLING 2016 Organizing Committee.

Miwa, M. and S. Ananiadou (2013). NaCTeM EventMine for BioNLP 2013 CG and PC tasks. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 94–98. Association for Computational Linguistics.

Miwa, M., R. Sætre, J.-D. Kim, and J. Tsujii (2010). Event Extraction with Complex Event Classification Using Rich Features. *Journal of Bioinformatics and Computational Biology*, **8**(1), pp. 131–146. DOI: 10.1142/S0219720010004586.

Miwa, M., P. Thompson, and S. Ananiadou (2012). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, **28**(13), pp. 1759–1765. DOI: 10.1093/bioinformatics/bts237.

Mykowiecka, A., M. Marciniak, and A. Kupść (2009). Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, **42**(5), pp. 923–36. ISSN 1532-0464. DOI: 10.1016/j.jbi.2009.07.007.

Nédellec, C., R. Bossy, J.-D. Kim, J.-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum (2013). *Proceedings of the BioNLP Shared Task 2013 Workshop*, chapter Overview of BioNLP Shared Task 2013, pp. 1–7. Association for Computational Linguistics.

Ohta, T., S. Pyysalo, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, S. Ananiadou, and J. Tsujii (2013). Overview of the Pathway Curation (PC) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 67–75. Association for Computational Linguistics.

Oosterhaven, J. (2015). Too many journals? Towards a theory of repeated rejections and ultimate acceptance. volume 103, pp. 261–265. ISSN 1588-2861. DOI: 10.1007/s11192-015-1527-4.

Pautasso, M. (2012). Publication Growth in Biological Sub-Fields: Patterns, Predictability and Sustainability. *Sustainability*, **4**(12), pp. 3234–3247. ISSN 2071-1050. DOI: 10.3390/su4123234.

Pennington, J., R. Socher, and C. Manning (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics. DOI: 10.3115/v1/D14-1162.

Peters, S. E., C. Zhang, M. Livny, and C. Ré (2014). A Machine Reading System for Assembling Synthetic Paleontological Databases. *PLOS ONE*, **9**(12), pp. 1–22. DOI: 10.1371/journal.pone.0113523.

Pyysalo, S., T. Ohta, M. Miwa, and J. Tsujii (2011). Towards Exhaustive Event Extraction for Protein Modifications. In *Proceedings of BioNLP 2011 Workshop*, pp. 114–123. Association for Computational Linguistics.

Ramshaw, L. and M. Marcus (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Řehůřek, R. and P. Sojka (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ELRA, Valletta, Malta.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. Macmillan. ISBN 0486240045.

Riaz, M. and R. Girju (2013). *Proceedings of the SIGDIAL 2013 Conference*, chapter Toward a Better Understanding of Causality between Verbal Events: Extraction and Analysis of the Causal Power of Verb-Verb Associations, pp. 21–30. Association for Computational Linguistics.

Riloff, E. (1996). Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pp. 1044–1049. AAAI Press. ISBN 0-262-51091-X.

Riloff, E. and J. Wiebe (2003). Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Sebastian, Y., E.-G. Siew, and S. O. Orimaye (2017). Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review*, **32**, p. e12. DOI: 10.1017/S0269888917000042.

Silver, D. L. (1996). The Parallel Transfer of Task Knowledge Using Dynamic Learning Rates Based on a Measure of Relatedness. *Connection Science*, **8**(2), pp. 277–294. DOI: 10.1080/095400996116929.

Smalheiser, N. R. and D. R. Swanson (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, **57**(3), pp. 149 – 153. ISSN 0169-2607. DOI: https://doi.org/10.1016/S0169-2607(98)00033-9.

Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, **34**(1), pp. 233–272. ISSN 1573-0565. DOI: 10.1023/A:1007562322031.

Spranger, M., S. Palaniappan, and S. Gosh (2016). Measuring the State of the Art of Automated Pathway Curation Using Graph Algorithms - A Case Study of the mTOR Pathway. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 119–127. Association for Computational Linguistics. DOI: 10.18653/v1/W16-2916.

Steeg, G. V. and A. Galstyan (2014). Discovering Structure in High-dimensional Data Through Correlation Explanation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 577–585. MIT Press, Cambridge, MA, USA.

Surdeanu, M., M. Valenzuela-Escárcega, G. Hahn-Powell, P. Jansen, D. Fried, D. Bell, and T. Hicks (2014). processors. `https://github.com/clulab/processors`.

Swanson, D. R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, **30**(1), pp. 7–18.

Swanson, D. R. (1986b). Undiscovered public knowledge. *The Library Quarterly: Information, Community, Policy*, **56**(2), pp. 103–118.

Swanson, D. R. and N. R. Smalheiser (1996). Undiscovered Public Knowledge: A Ten-year Update. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 295–298. AAAI Press.

Taylor, A., M. Marcus, and B. Santorini (2003). *The Penn Treebank: An Overview*, pp. 5–22. Springer Netherlands, Dordrecht. ISBN 978-94-010-0201-1. DOI: 10.1007/978-94-010-0201-1_1.

Thrun, S. and J. O'Sullivan (1995). Clustering Learning Tasks and the Selective Cross-Task Transfer of Knowledge. Technical report. DOI: 10.1007/978-1-4615-5529-2_10.

Toutanova, K. and D. Chen (2015). Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66. Association for Computational Linguistics. DOI: 10.18653/v1/W15-4007.

Toutanova, K., D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon (2015). Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509. Association for Computational Linguistics. DOI: 10.18653/v1/D15-1174.

Tsai, C.-T. and D. Roth (2016). Concept Grounding to Multiple Knowledge Bases via Indirect Supervision. *Transactions of the Association for Computational Linguistics*, **4**, pp. 141–154. ISSN 2307-387X.

Valenzuela-Escárcega, M., G. Hahn-Powell, D. Bell, T. Hicks, E. Noriega, and M. Surdeanu (2015a). REACH. `https://github.com/clulab/reach`.

Valenzuela-Escárcega, M. A., Ö. Babur, G. Hahn-Powell, D. Bell, T. Hicks, E. Noriega-Atala, X. Wang, M. Surdeanu, E. Demir, and C. T. Morrison (2017). Large-scale automated reading with Reach discovers new cancer driving mechanisms. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*, pp. 201–203.

Valenzuela-Escárcega, M. A., G. Hahn-Powell, D. Bell, and M. Surdeanu (2016). SnapToGrid: From Statistical to Interpretable Models for Biomedical Information Extraction. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 56–65. Association for Computational Linguistics. DOI: 10.18653/v1/W16-2907.

Valenzuela-Escárcega, M. A., G. Hahn-Powell, T. Hicks, and M. Surdeanu (2015b). A Domain-independent Rule-based Framework for Event Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Software Demonstrations*, pp. 127–132. ACL-IJCNLP 2015. DOI: 10.3115/v1/P15-4022.

Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2015c). Description of the Odin Event Extraction Framework and Rule Language.

Valenzuela-Escárcega, M. A., G. Hahn-Powell, and M. Surdeanu (2016). Odin's Runes: A Rule Language for Information Extraction. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. LREC.

Venugopal, D., C. Chen, V. Gogate, and V. Ng (2014). Relieving the Computational Bottleneck: Joint Inference for Event Extraction with High-Dimensional Features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 831–843. Association for Computational Linguistics. DOI: 10.3115/v1/D14-1090.

Wu, F., R. Hoffmann, and D. S. Weld (2008). Information Extraction from Wikipedia: Moving Down the Long Tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 731–739. ACM. ISBN 978-1-60558-193-4. DOI: 10.1145/1401890.1401978.

Yakushiji, A., Y. Miyao, T. Ohta, Y. Tateisi, and J. Tsujii (2006). Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 284–292. Association for Computational Linguistics.

Yang, J., J. McAuley, and J. Leskovec (2014). Detecting Cohesive and 2-mode Communities Indirected and Undirected Networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pp. 323–332. ACM, New York, NY, USA. ISBN 978-1-4503-2351-2. DOI: 10.1145/2556195.2556243.

Zhang, C. (2015). *DeepDive: a data management system for automatic knowledge base construction*. Ph.D. thesis, University of Wisconsin-Madison.