Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph

Gus Hahn-Powell **Marco Valenzuela-Escárcega** **Mihai Surdeanu**
University of Arizona, Tucson, AZ, USA
{hahnpowell, marcov, msurdeanu}@email.arizona.edu

Abstract

We introduce a modular approach for literature-based discovery consisting 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. Our 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 Introduction

Since at least 1990, there has been exponential growth in the number of academic papers published annually in the biomedical domain (Pautasso, 2012). For example, the number of English language biomedical publications indexed by PubMed[1] alone since 1900 has now surpassed 25 million[2]. Over 17 million of these were published between 1990 and 2015.

Although a number of information extraction (IE) 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 (1986) described this problem as "undiscovered public knowledge".

In this work, we propose 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 (Smalheiser and Swanson, 1998; Bekhuis, 2006), 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.[3] 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).

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

[2] https://www.ncbi.nlm.nih.gov/pubmed/?term=%221900%22%5BPDAT%5D%20%3A%20%222017%22%5BPDAT%5D&cmd=DetailsSearch

[3] In other words, we trust that the authors' statements are correct. Although these statements often contain causal language (e.g., "A causes B"), we avoid referring to these relations as causal, since our approach does not attempt to verify quantitative findings directly.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 103–108
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-4018

2 Previous Work

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). All of these systems read individual facts, which makes Swanson's observation even more valid today. Attempts have been made in the biomedical domain to assemble relations extracted by machine reading into coherent models (Hahn-Powell et al., 2016). This domain is known for the complexity of its models (Lander, 2010), which has spurred research on improved visualizations for actionable insights (Dang et al., 2015).

Our 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.

3 Approach

Our 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. In this section, we describe our machine reading and assembly approach. In Section 4 we describe the search engine over this influence graph.

We introduce methods for machine reading and assembly in 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 we discuss 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.

3.1 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)[4]. 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 IE framework (Valenzuela-Escarcega et al., 2016) 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 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 1, and described below.

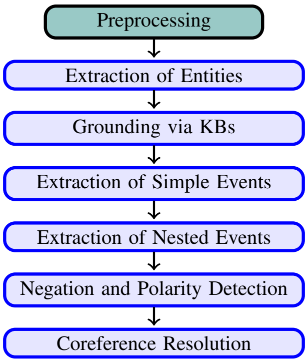

Preprocessing
Extraction of Entities
Grounding via KBs
Extraction of Simple Events
Extraction of Nested Events
Negation and Polarity Detection
Coreference Resolution

Figure 1: The REACH pipeline, which includes detection of hedging and negation. In addition to the components shown, REACH detects the scope of biological contexts such as cell line, tissue type, and species (not covered here).

Entity Extraction and Resolution

Because of its focus on a single domain (molecular biology), REACH leverages domain-specific resources such as curated knowledge bases (e.g., Uniprot[5] for protein names, and PubChem[6] for other chemicals) to perform entity recognition and resolution.

REACH's named entity recognizer (NER) is a hybrid model that combines a rule-based component with a statistical one.[7] The output of the NER system is then matched against the set of knowledge bases in order to "ground" these textual mentions to real-world entities through ID assignment. This grounding component enables the deduplication of entity nodes during graph construction (the assembly phase). That is, all synonyms of the

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

[5] `http://www.uniprot.org`

[6] `https://pubchem.ncbi.nlm.nih.gov`

[7] The details of this hybrid NER architecture are omited for brevity.

same protein are mapped to the same node in the influence graph.

Event Extraction

REACH uses a two-stage bottom-up strategy for event extraction, following biochemical semantics inspired by BioPAX (Demir et al., 2010). First, we identify biochemical reactions that operate directly on entities, initially ignoring their catalysts and other controllers, e.g., phosphorylation of a protein. We call these events "simple". REACH uses a domain-specific taxonomy for specifying the selectional restrictions on event arguments. Next, nested events are captured. For example, during this stage, catalysts that control simple events such as phosphorylations are extracted. During graph construction, these nested events are "flattened" into binary influence links. For example, the nested interaction `PositiveRegulation(Controller:A, Controlled:Phosphorylation(B))` is reduced to A $\rightarrow$ B, where the arrow indicates a left-to-right influence relation. Additionally, these links preserve the type (e.g., phosphorylation) and polarity of the biochemical interaction, e.g., a positive regulation is equivalent to a "promotes" link, whereas a negative regulation reduces to "inhibits".[8]

Coreference Resolution and Negation

The coreference resolution component in REACH adapts the algorithm of Lee et al. (2013) to the biomedical domain, where it operates both over entity mentions (e.g., by resolving pronominal and nominal mentions such as "it" or "this protein" to the corresponding entity) and event mentions (e.g., "this interaction" is resolved to an actual event) (Bell et al., 2016).

The negation detection module identifies explicit statements that a reaction does not occur (e.g., "ZAP70 does not induce TRIM phosphorylation") in a particular experimental context. REACH also handles more subtle linguistic phenomena such as reversing the polarity of events. For example, the naive interpretation of the text "decreased PTPN13 expression enhances EphrinB1 phosphorylation" yields a positive regulation (due to "enhances"). The polarity correction module changes this to a negative regulation due to the presence of the "decreased" modifier.

3.2 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 our 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.

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[9] 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.

[8] Polarity detection is more complicated in practice, because some simple events have polarity information as well, which needs to be taken into account when flattening events. We ignore this situation here for brevity.

[9] Function words are ignored.

4 Influence Graph Search Engine

Conventional visualization of any sufficiently complex network suffers from the "hairball" (Lander, 2010) problem. To mitigate this obfuscation, we introduce a search user interface (UI)[10] 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 2 for a detailed description of the UI).

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 4.1.

4.1 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. We provide several scores for each edge, which 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. IDF scores were calculated 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.

4.2 Influence Graphs

In Figure 3, we show an example of output for our domain-specific IE system (see §3.1). The output demonstrates the system's ability to capture chains of biomolecular interactions. Output for our open domain IE system (see §3.2) is shown in Figure 4, which showcases the system's ability to uncover indirect links between concepts across domains.

5 Evaluation

An early version 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 query against the open access subset of PubMed to identify papers relevant to her use case. We then extracted influence relations from the retrieved documents and presented the ranked results to the biologist.

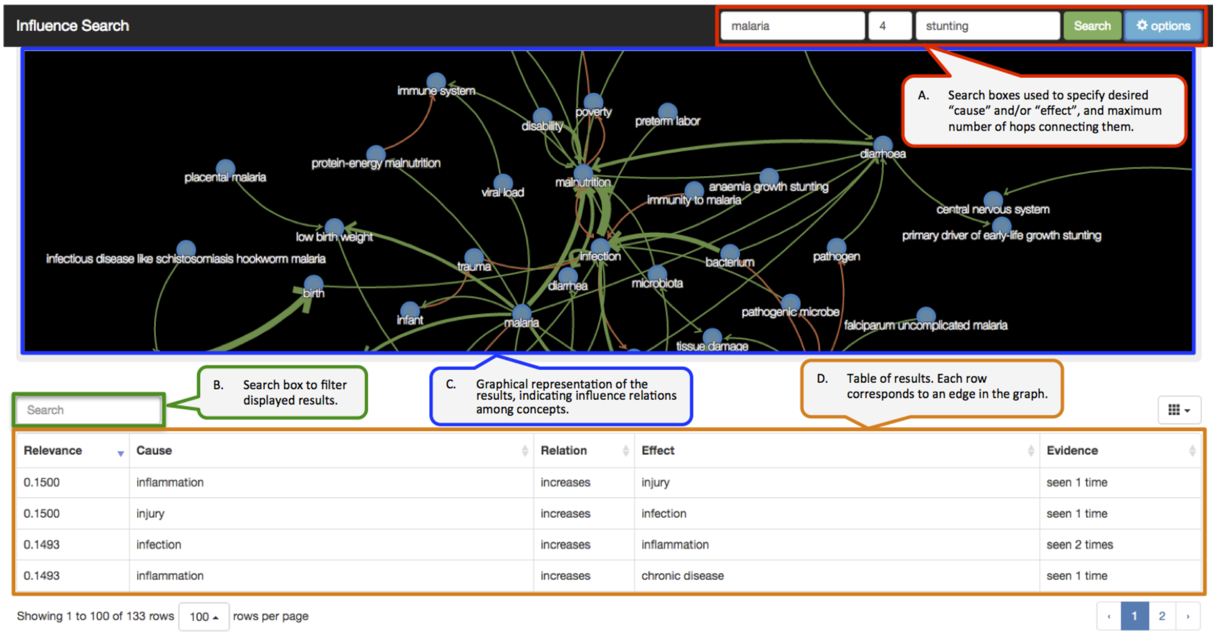
Through her exploration, the biologist refined her search terms of interest and the process was repeated for three cycles over two days. The result was a model containing 35 core concepts such as "EBF", "improved water access", and "government subsidy", and 48 directed influence relations holding between 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, remarking that "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." We consider this success as a first step toward bridging research islands through Swanson linking.

6 Conclusion

We introduced a search engine that operates over influence graphs that were automatically mined from scientific literature. Our 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

[10] More information on this system can be found at http://clulab.cs.arizona.edu/demos/influence-search



A. Search boxes used to specify desired "cause" and/or "effect", and maximum number of hops connecting them.

B. Search box to filter displayed results.

C. Graphical representation of the results, indicating influence relations among concepts.

D. Table of results. Each row corresponds to an edge in the graph.

Relevance	Cause	Relation	Effect	Evidence
0.1500	inflammation	increases	injury	seen 1 time
0.1500	injury	increases	infection	seen 1 time
0.1493	infection	increases	inflammation	seen 2 times
0.1493	inflammation	increases	chronic disease	seen 1 time

Showing 1 to 100 of 133 rows 100 rows per page

Figure 2: 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. **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 "seen X times" entry corresponding to that row.

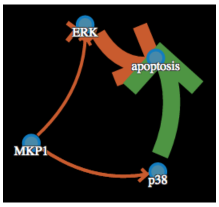


Figure 3: One result of *MKP1*'s role in *apoptosis*, which demonstrates multiple pathways with competing downstream effects. The query for this result was constrained to a maximum distance of three intervening nodes.

that this deep reading and search approach reduces the effort needed to uncover "undiscovered public knowledge" (Swanson, 1986). With the aid of this tool, a domain expert reduced her model building time from months to two days.

In future work, we will strengthen our assembly approach by improving node and edge deduplication. This 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. Additionally, we plan to add a context filter to the search fields, which will allow the user to focus results by context, e.g., "show results just for *pancreatic cancer*", or "show impact factors to children's health in *Sudan*."

Acknowledgments

This work was funded by the DARPA Big Mechanism program under ARO contract W911NF-14-1-0395 and by the Bill and Melinda Gates Foundation HBGDki Initiative. The authors declare a financial interest in lum.ai, which licenses the intellectual property involved in this research. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007.

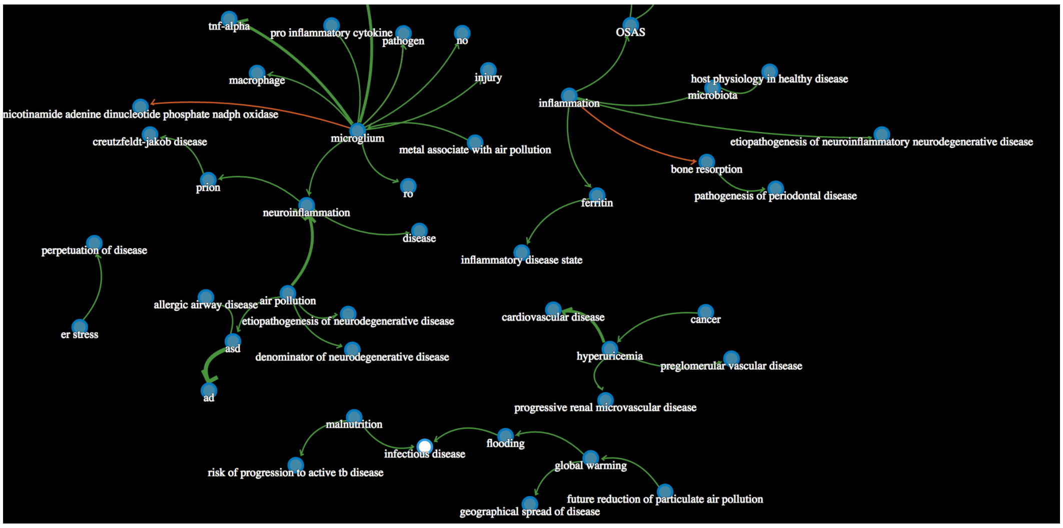

tnf-alpha
pro inflammatory cytokine
pathogen
no
OSAS
macrophage
injury
host physiology in healthy disease
microbiota
inflammation
nicotinamide adenine dinucleotide phosphate nadph oxidase
creutzfeldt-jakob disease
microglium
metal associate with air pollution
etiopathogenesis of neuroinflammatory neurodegenerative disease
bone resorption
prion
ro
pathogenesis of periodontal disease
ferritin
neuroinflammation
disease
perpetuation of disease
inflammatory disease state
allergic airway disease
air pollution
etiopathogenesis of neurodegenerative disease
cardiovascular disease
cancer
er stress
asd
denominator of neurodegenerative disease
hyperuricemia
preglomerular vascular disease
ad
progressive renal microvascular disease
malnutrition
infectious disease
flooding
risk of progression to active tb disease
global warming
future reduction of particulate air pollution
geographical spread of disease

Figure 4: A sample of the top 50 indirect connections between *pollution* and *disease*. In this example, the connection is constrained to at most three intervening nodes between these two concepts.

Open information extraction from the web. In *IJCAI*. volume 7, pages 2670–2676.

Tanja B Bekhuis. 2006. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Library* 3(2).

Dane Bell, Gus Hahn-Powell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2016. Sieve-based coreference resolution in the biomedical domain. In *LREC*.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on BioNLP: Shared Task*. ACL, pages 10–18.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*. volume 5, page 3.

Tuan Nhon Dang, Paul Murray, and Angus Graeme Forbes. 2015. Pathwaymatrix: Visualizing binary relationships between proteins in biological pathways. In *BMC Proceedings*. BioMed Central Ltd, volume 9, page S3.

Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'eustachio, Carl Schaefer, Joanne Luciano, et al. 2010. The biopax community standard for pathway data sharing. *Nature biotechnology* 28(9):935–942.

Gus Hahn-Powell, Dane Bell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2016. This before that: Causal precedence in the biomedical domain. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*.

Arthur D Lander. 2010. The edges of understanding. *BMC biology* 8(1):40.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4).

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. *Proceedings of the BioNLP Shared Task 2013 Workshop*, Association for Computational Linguistics, chapter Overview of BioNLP Shared Task 2013, pages 1–7. http://aclweb.org/anthology/W13-2001.

Marco Pautasso. 2012. Publication growth in biological sub-fields: patterns, predictability and sustainability. *Sustainability* 4(12):3234–3247.

Shanan E. Peters, Ce Zhang, Miron Livny, and Christopher Ré. 2014. A machine reading system for assembling synthetic paleontological databases. *PLOS ONE* 9(12):1–22. https://doi.org/10.1371/journal.pone.0113523.

Neil R Smalheiser and Don R Swanson. 1998. Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer methods and programs in biomedicine* 57(3):149–153.

Don R Swanson. 1986. Undiscovered public knowledge. *The Library Quarterly* 56(2):103–118.

Marco A Valenzuela-Escarcega, Gustave Hahn-Powell, and Mihai Surdeanu. 2016. Odin's runes: A rule language for information extraction. In *LREC*.

Ce Zhang. 2015. *DeepDive: a data management system for automatic knowledge base construction*. Ph.D. thesis, Citeseer.