Stony Brook University

Knowledge Extraction from Diverse Biomedical Corpora with Applications in Healthcare: Bridging the Translational Gap

A Dissertation Presented
By
Ritwik Banerjee

to
The Graduate School
in Partial Fulfullment of the Requirements
for the Degree of
Doctor of Philosophy
in
Computer Science
Stony Brook University

December, 2015

Stony Brook University

The Graduate School

Ritwik Banerjee

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Yejin Choi, Advisor
Assistant Professor, Department of Computer Science

Paul Fodor, Chairperson of Defense
Research Assistant Professor, Department of Computer Science

I. V. Ramakrishnan, Co-Advisor
Professor, Department of Computer Science

Niranjan Balasubramanian
Research Assistant Professor, Department of Computer Science

Hasan Davulcu, External Committee Member
Associate Professor, Department of Computer Science and Engineering
Arizona State University

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Knowledge Extraction from Diverse Biomedical Corpora with Applications in Healthcare: Bridging the Translational Gap

by

Ritwik Banerjee

Doctor of Philosophy

in

Computer Science

Stony Brook University

2015

A wealth of knowledge in the biomedical domain is available in unstructured or semi-structured data repositories as natural language narratives. Much of this knowledge can provide immediate and tangible benefits in patient welfare and the healthcare industry. Extracting relevant knowledge from these natural language sources and providing them as structured information suitable for immediate real-time consumption in clinical settings is, however, a manual process restricted to human domain experts. As a result, it is expensive and time-consuming. A very real consequence of this is that the *journey* made by medical "knowledge nuggets" from research publications to patient care settings like hospitals often take several years. Even so, the knowledge still gets presented to clinicians in natural language – unsuitable for machine consumption, and an impediment to the pace of work often demanded of clinicians (*e.g.* in emergency rooms).

Automatic extraction of this knowledge is a challenging task. Biomedical research literature is replete with language constructs that are highly specific to not just the domain, but internal sub-domains. The linguistic semantics used in discussions of, say, diabetes, are very different from the semantics used to discuss diseases like malaria that are caused by external agents. Moreover, being research literature, authors typically write for readers with a fair amount of *encyclopaedic* domain knowledge. Consequently, important information can often only be gleaned by identifying causal relations that are *implicit*. Standard information extraction methods that depend on identifying causality in text usually require explicit discourse connectives like "because", "since", etc. Additionally, they manage to extract only those relations that are expressed within the span of a single sentence.

This proposal presents a novel relation learning methodology for biomedical natural language that is able to infer relations where (a) the relation is *implicit*, and (b) the related entities do not co-occur within the span of a single sentence. We show that our technique outperforms a sentence-level supervised classification approach. Further, as a human-in-the-loop (HITL) model, it is capable of augmenting biomedical knowledge bases quickly and accurately. Finally, we contribute two novel applications that demonstrate the use of such relational knowledge in providing *real-time* clinical decision support.

Contents

Acknowledgements

I would like to thank my advisor, Professor Yejin Choi, for her continuous support and guidance for the past five years. She has been an exceptional mentor, and has repeatedly enthused me with the desire to explore deeper into the realm of open research. Her help extends much beyond just this thesis, though. I have often obtained valuable advice about various aspects of research in general, and also about career. Her professionalism and hard work are infectious! I hope to attain the same dedication to research as she does. I would also like to thank my co-advisor, Professor I. V. Ramakrishnan, who has provided me with valuable support and guidance about research direction and career. He is the primary reason why I got interested in the biomedical domain within natural language processing as a whole. His practical outlook on research objectives has taught me some significant lessons in the last few years, and I am certain that these lessons will continue to be of paramount importance in the future. At times, he has also gone out of his way to help me with non-academic issues. I am indebted to such amazing mentors.

I am also very grateful to Professor Niranjan Balasubramanian. His collaboration has been invaluable. I truly admire his hands-on approach to research. No matter how busy he is, he is always full of enthusiasm and never shies away from getting his hands dirty with raw data! Without his help, an important component of this thesis would have probably taken much longer to mature. I would like to take this opportunity to also thank Dr. Mark Henry, professor and chair of Emergency Medicine at Stony Brook University Hospital. His support and insights from the biomedical perspective have not only helped me understand the research domain in broader context, but also paved the way for multiple future research prospects. I would also like to thank my other committee members, Professors Paul Fodor and Hasan Davulcu. They took valuable time out of their busy schedules to share with me their advice and suggestions.

In addition to the support and guidance I received from the academic environment, I would like to thank those who provided the emotional support I have often needed over the last few years. Such a support network is crucial for PhD students to stay focused through their long and arduous journey. To begin with, I want to thank my fiancee Stephanie Brown for her love and support. Her presence by my side was often the catalyst I needed to keep going even during the most difficult moments. Lastly, I want to thank those who are responsible for this journey even before it began: my family – my parents in particular. Their hard work spanning over decades, their countless sacrifices for my sister and myself, their unconditional love and guidance, and many other aspects for which I can find no words. Without them, I would not be the person I am today; I would not have had the perseverance or the passion for research.

Introduction

The biomedical sciences, at least in terms of sheer volume of new information, have recently been experiencing explosive growth (Druss and Marcus, 2005; Boissier, 2013). New results are being published in the research literature in this domain with increasing frequency. For instance, PubMed, the most widely used repository of biomedical articles, has tripled its growth rate over the last decade (Andronis et al., 2011). Much of this information can lead to immediate and tangible benefits in the healthcare industry, and decidedly, in patient welfare. One of the most important types of knowledge that can directly provide these benefits is *relational information* connecting behavioral, social and biomedical entities or concepts. These include the relation between a lifestyle choice and a disease (*e.g.* smoking and lung cancer), a drug and a symptom (*e.g.* Tylenol and headache), etc. In addition to well established relations like these two examples, new information is continuously being produced and made available in biomedical research literature. Automatic extraction of actionable information from these sources, however, remains highly challenging not only because it is presented in natural language, but also because the language used in these publications is highly specialized. In fact, the language is specific to not just the domain as a whole, but sub-domains within biomedicine. For instance, the linguistic semantics used in literature on psychological diseases are very different from the semantics used in discussions of infectious diseases. Moreover, since the narratives are intended for highly knowledgeable readers, distilling the knowledge presented in this data requires a lot of *encyclopaedic* domain knowledge.

In spite of the challenges, extracting relevant information from research literature cannot be left as a manual process to be carried out by human domain experts. This is inherently expensive due to the associated labor cost. Moreover, it leads to unacceptable delays in updating structured knowledge bases (KBs), and is prone to errors of omission. Drug interaction databases, for example, are seldom complete because their update periods can be as long as three years (Rodriguez-Terol et al., 2009). It has also been noted that databases on adverse drug effects miss nearly a quarter of all clinical trials reporting such effects (Derry, Loke, and Aronson, 2001). A medical "knowledge nugget" from research literature can thus take a long time before it is incorporated in databases used by hospitals and other healthcare facilities. These delays can have catastrophic consequences, as illustrated by the unfortunate incident of over $27,000$ heart attacks caused by the pain relief drug *rofecoxib* before the drug was recalled in 2004[1]. Many other drugs have had similar results in the past. Examples include *cisapride*, withdrawn after 80 reported fatalities[2] and *cerivastatin*, recalled after 52 deaths and nearly 400 cases of hospitalization[3]. The need for scalable learning is

[1] MSNBC Staff, "Report: Vioxx Linked to Thousands of Deaths", www.nbcnews.com, Oct. 6, 2004
[2] FDA, "Propulsid (cisapride) Dear Healthcare Professional Letter Jan 2000", www.fda.gov, Jan 24, 2000.
[3] Furberg and Pitt, 2001.

not just limited to adverse effects. Extraction of therapeutic relations between drugs and diseases from natural language to create structured knowledge is of vital importance. Such knowledge has important real-world applications including healthcare information retrieval (Hanbury, 2012; Lialiou and Mantas, 2014), bioinformatics research (Tatonetti et al., 2012; Li and Lu, 2013) and clinical decision support (Duke and Friedlin, 2010; Banerjee et al., 2014; Banerjee et al., 2015).

This gap between research and clinical applications has been discussed by biomedical researchers and healthcare practitioners for over a decade, with a number of publications commenting on how we need to *translate* knowledge from research to clinical settings (*e.g.* Lenfant, 2003; Glasgow and Emmons, 2007; Laan and Boenink, 2012; Burnand, 2015), thus giving rise to the phrase "translational research" in the biomedical domain. Several authors have pointed out that the bulk of translational research so far has focused on bridging the gap between basic research and clinical investigations, but not on taking the results of clinical investigations and translating them into evidence-based healthcare practice (*e.g.* Lenfant, 2003; Clyne et al., 2014).

Accurate extraction of relational information from biomedical literature is an important step in this *translation pipeline*. In order to have utility in clinical settings, this step needs to be *fast* enough to make research findings available with little or no delay, and *accurate* so that the extracted knowledge can, indeed, have a positive impact on patient care. Current relation extraction techniques can, with reasonable accuracy, extract relations explicitly mentioned within the span of a single sentence. In biomedical research publications, however, we observed that a majority of the relations are not expressed within a single sentence. Further, the discourse is often *implicit*. These aspects make standard relation extraction approaches unsuitable for this domain.

The first part of this thesis outlines our contribution to the first half of the translation pipeline: a novel relation learning methodology that combines elements of text-based relation extraction techniques and statistical relation learning in knowledge graphs. This, effectively, allows for *relation inference* in the absence of clear discourse in texts and in the absence of connecting paths in knowledge graphs.

We show that this approach is a significant improvement over a supervised classification baseline, and is able to distill relational knowledge even when (a) **the related entities are not in the same sentence**, and (b) **there is no explicit discourse connective**. This has a twofold advantage over traditional relation extraction frameworks. First, it enables inference based on simple lexico-syntactic constructs even when the underlying text consists of long and complex sentences. Second, by inferring relations that are beyond the scope of sentence-level extraction systems, it greatly reduces the omission errors that currently afflict medical KBs (Derry, Loke, and Aronson, 2001; Reaume, 2012). Being fully computational, it is obviously a much faster process than any manual curation framework. In addition, the high precision and yield indicate that our methodology can be adapted into an *expert-in-the-loop* process that can very quickly provide a completely accurate list of biomedical relations. In this manner, we have obtained therapeutic relations between drugs and diseases/symptoms, and established that this can be used to augment structured KBs like the Unified Medical Language System (UMLS) (Lindberg, Humphreys, and McCray, 1993) and DrugBank (Knox et al., 2011).

It is worth noting that in spite of being indispensable, translational research is not by any means a complete substitute for human domain expertise. Manually curated methods have produced a sizable body of structured and *semi-structured* KBs that provide accurate information about thousands of medical entities. From the perspective of healthcare applications in clinical settings, which constitute the later stages of the translation pipeline, these KBs are extremely important. In fact, many are used as standard reference points by practicing clinicians. However, in addition to these being incomplete, this practice has several other pitfalls. Note that patient care is not

just a single instance of drug use for treatment. It consists of a dynamic, and often iterative (especially in disease management), process, which demands consolidating many different *types* of knowledge. For example, if a patient is exhibiting certain symptoms, s/he may be asked to undergo a laboratory test. Depending on the test results, the physician may prescribe a drug, alter the dosage, suggest lifestyle changes, ask for more laboratory tests, etc. This requires integrating information across heterogeneous KBs that have been developed for divergent uses and different types of users (*e.g.* radiologists, pharmacists, patients).

The second part of our work pertains to integrating such heterogeneous information sources for practicable patient care. Here, we present our contribution to the second half of the translation pipeline: bringing the outcome of biomedical research investigations into clinical practice. We describe two novel applications of medical information extraction in clinical decision support. In both applications, our focus is on prevention of adverse drug events (ADEs) – events that are unintended and undesired reactions experienced by an individual due to use, misuse or discontinuation of medication. Several studies have reported that among the adult population, a high percentage of emergency room (ER) visits are caused by ADEs (Zed et al., 2008; Jayarama, Shiju, and Prabahakar, 2012). As there exist way too many drugs, physicians cannot be expected to remember all possible ADEs associated with them. Semi-structured drug data repositories help to some extent by providing a list of conditions associated with the adverse effects of a drug. Not all adverse effects manifest as observable symptoms, however. Some can only be confirmed by laboratory tests. But, even though laboratory testing is the single highest-volume medical activity driving clinical decision making, information pertaining to them is not integrated with drug KBs. As a result, the process of ordering diagnostic tests and acting upon them remains vulnerable to errors (Singh, 2013; Zhi et al., 2013). Our first application is a system that carefully harnesses patient electronic health record (EHR) with a collection of different *types* of medical data repositories to automatically suggest laboratory tests to confirm (or invalidate) potential adverse effects of a patient's drug regimen. The performance of this system is measured using *sensitivity*, which indicates its ability to correctly identify the laboratory test required to confirm ADEs.

Our second application presents a patient centered approach toward identification and attribution of ADEs. As mentioned earlier, information about ADEs is often available in online drug KBs in the form of narrative texts, and serves as the physician's primary reference point for ADE attribution and diagnosis. Manually reviewing these narratives, however, is an error prone and time consuming process, especially due to the prevalence of polypharmacy. So ER health care providers, especially given the heavy volume of traffic, often either skip this step or at best do it rather perfunctorily. This causes ADEs to be missed or misdiagnosed, often leading to extensive and unnecessary testing and treatment, including hospitalization. This part of the thesis describes a system that automates the detection of ADEs and provides a list of suspect drugs, ranked by their likelihood of causing the patient's complaints and symptoms. The input data, i.e., medications and complaints, are obtained from triage notes that often contain descriptive language. Our application utilizes heterogeneous information sources (including drug KBs) to refine and transform these descriptions as well as the online database narratives using a natural language processing (NLP) pipeline. Our work incorporates a domain-specific entity normalization method. We then employ ranking measures to establish correspondence between the complaints and the medications. Our evaluation based on actual cases demonstrates that this system achieves high precision and recall.

The body of this thesis constitutes three chapters, describing our contribution to the translation pipeline of bringing research knowledge into a structured, actionable form that can be directly applied in clinical settings. Chapter 1 presents the details of our relation learning methodology from research literature. Chapters 2 and 3 describe the two applications introduced above.

Chapter 1

Relation Inference in Biomedical Texts

The large amount of literature in biomedical science and structured knowledge bases (KBs) available today makes it difficult for researchers and medical practitioners alike to absorb and retain all the information relevant in their fields. The problem is particularly acute for physicians working in patient care settings, who must often tend to heavy traffic, and therefore cannot afford the luxury of manually perusing research literature to stay abreast of the latest findings. From the vantage point of healthcare providers, *relational information* is extremely important. Since much of patient care relies on understanding the relation between drugs, diseases, symptoms, lifestyle aspects, etc., there is a growing demand for structured information extracted from literature (Ananiadou and Mcnaught, 2005).

The discovery of such knowledge usually begins in research settings, goes through rigorous clinical investigations and finally reaches the proverbial "bedside", where the knowledge is finally applied for the benefit of patients. The process of translating basic scientific findings into therapeutic interventions for patients has often been called the "bench-to-bedside" process (*e.g.* Zerhouni (2005) and Luciano et al. (2011)), or *translational* research (*e.g.* Birmingham (2002), Woolf (2008), and Butler (2008)). As described in the introduction, there remains a significant gap between the research findings and their application in healthcare. This gap is both temporal – resulting in significant delay between a discovery and its use in healthcare (Rodriguez-Terol et al., 2009) – and quantitative, *i.e.* a large proportion of research knowledge does not reach the realm of patient care (Derry, Loke, and Aronson, 2001). In an attempt to significantly reduce this translational gap, the current chapter describes our contribution toward extracting relational knowledge from biomedical research literature. In particular, we focus on relations that can be broadly categorized as *beneficial* or *harmful*, and present a novel methodology to infer such relations from research narratives.

1.1 Domain Characteristics

Any natural language processing task is burdened by the ambiguity and variability of human language, and biomedical research literature is replete with both. Biomedical literature is considered to be one of the most difficult domains for NLP tasks, and several authors have used various metrics to support this intuition (*e.g.* Zeng-Treitler et al. (2007), Leroy et al. (2008), and Wu et al. (2015)). Moreover, it is characterized by highly specialized *sub-domains* providing very different perspec-

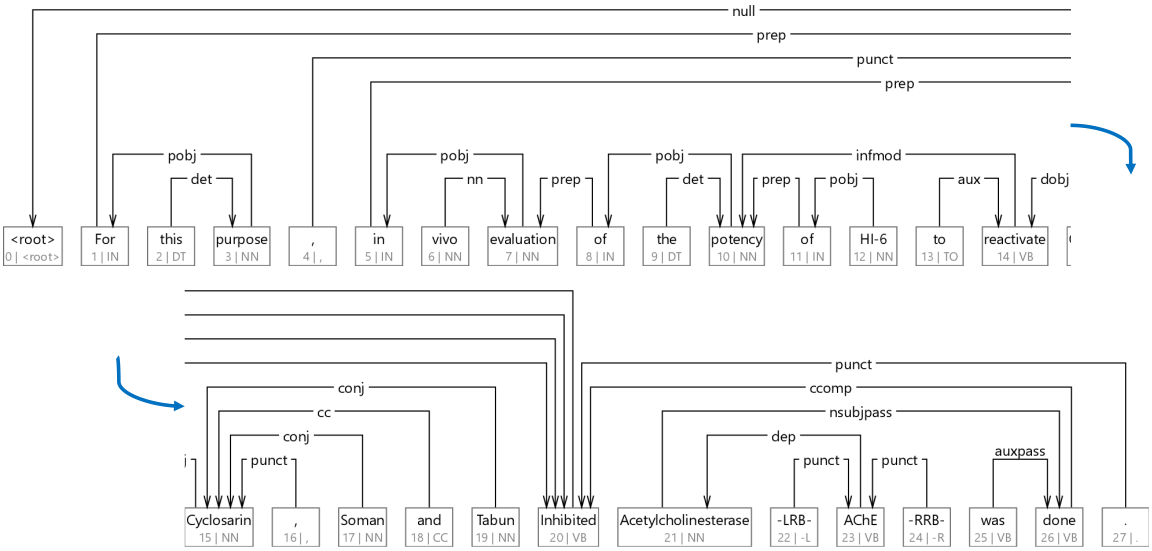

Figure 1.1: The syntactic dependency parse tree, obtained using TurboParser v2.2[1] for the sentence "For this purpose, in vivo evaluation of the potency of HI-6 to reactivate Cyclosarin, Soman and Tabun Inhibited Acetylcholinesterase (AChE) was done." Errors often include wrong POS tags. The acronym 'AChE' and 'Inhibited' in a compound noun are both tagged as verbs while 'in' in *in vivo* is tagged as a preposition. These in turn lead to parsing errors.

tives, often for similar issues. This degree of specialization in scientific endeavors is reflected in the use of specialized language. Its drawback, however, is the creation of what has been called "islands of knowledge" (Andronis et al., 2011) that mystify the interconnections between these sub-domains (Weeber et al., 2000). As a result, lexical synonymy and polysemy are exceedingly common (Marrero et al., 2012).

Authors also often condense a lot of information by using features such as long compound nouns, appositives and relative clauses. The sentences also often exhibit complex instances of right node raising. A combination of these characteristics render standard parsing techniques prone to errors. Relation extraction methods that depend on lexico-syntactic features (see Sec. 1.2) thus suffer from the resultant inaccuracies. The dependency parse tree of Fig. 1.1 presents a typical example of such errors. In many ways, the sentence used there is quite representative of the research literature domain. Sentences are often long and complex with multiple intervening entities. On a random sample representing 10% of the entire PubMed Central[2] (PMC) repository, we found that the average sentence length is 26 words (Wu et al. (2015) report an average sentence length of 26.1 words on another dataset that largely overlaps PMC). Further, the average distance between a drug and a disease mention was 8 words, with 28.9% of the relation bearing sentences having at least two drug mentions.

Additionally, a surprisingly small fraction of relations are expressed within the span of a single-sentence. We found that for more than 44% of the drug-disease relations present in the Unified Medical Language System[3] (UMLS) database, there were no sentences that contained both the drug and the disease. Clearly, sentence-level relation extraction techniques, even if they succeed in correctly parsing complicated domain-specific language, are bound to have low recall.

[1] Martins, Almeida, and Smith (2013)

[2] PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/) is an electronic archive of over 3.5 million articles. It was developed and is maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH).

[3] Lindberg, Humphreys, and McCray, 1993.

1.2 Related Work

A large body of work in information extraction and related areas focuses on learning relational knowledge. Much of the prior work in this direction can be put down as belonging to one of two broad categories: *relation extraction from texts*, and *relation learning from knowledge graphs*. In this section, we provide a review of these two approaches. In particular, we discuss, in some detail, statistical learning methods that employ latent variable models, and are pertinent to the approach adopted in this thesis.

1.2.1 Relation Extraction from Texts

Extracting relational information from unstructured natural language data has a multitude of applications, including information retrieval and natural language understanding, that require an understanding of the semantic relations between various entities. Widely studied semantic relations in the non-medical domain include relations between the three entity-types *person*, *organization* and *location*. For example, from sentences like "Stony Brook is one of Long Island's major centers of education." and "Bill Gates co-founded Microsoft Inc.", state-of-the-art techniques can extract relations of the type `located-in(Stony Brook, Long Island)` or `affiliated-with(Bill Gates, Microsoft)`. Even though relations may generally be defined between several entities, a majority of current research has focused on extracting binary relations that can be represented as triples of the form (e_i, r_k, e_j) where r_k is the relation, and e_i and e_j are the related entities. This problem has been explored in several domains by Agichtein and Gravano (2000), Mooney and Bunescu (2005), Xu, Uszkoreit, and Li (2007), Hoffmann et al. (2011), and Xu et al. (2013), among others. In the biomedical domain, much of prior research has focused on very specific relations such as protein-protein interactions (Yakushiji et al., 2006; Airola et al., 2008), gene-protein interactions (Fundel, Kuffner, and Zimmer, 2007), drug-drug interactions (Segura Bedmar, Martínez, and Herrero Zazo, 2013), drug-disease treatments (Xu and Wang, 2013) and adverse effects of single drugs (Liu and Chen, 2013; Nikfarjam et al., 2015).

Next, we discuss supervised approaches to relation extraction, delving into several feature- and kernel-based methods, and then a discussion of semi-supervised approaches. Finally, we briefly review distant supervision methods pertaining to relation extraction.

Supervised Methods

The most successful approaches have usually been supervised techniques using deep lexico-syntactic features (Culotta and Sorensen, 2004; Mooney and Bunescu, 2005; Kim et al., 2011; Nédellec et al., 2013; Segura Bedmar, Martínez, and Herrero Zazo, 2013), where the relation extraction task is formulated as a binary classification problem. This body of work can broadly be divided into *feature-based* and *kernel-based* methods. One advantage of the latter is that they enable polynomial-time exploration of large feature spaces. Of particular note in this area are the approaches using tree kernels (Zelenko, Aone, and Richardella, 2003), subsequence kernels (Mooney and Bunescu, 2005) and dependency kernels (Bunescu and Mooney, 2005). More recent approaches have shown that such kernels may be combined to further improve relation extraction results (Nguyen, Moschitti, and Riccardi, 2009).

The classification task is formulated in terms of a scoring function $f_k(S)$, where S is a sentence containing two entities e_i and e_j, and r_k is the relation under consideration. The scoring function typically has the range $[-1, 1]$, and positive/negative values determine whether or not the entities

e_i and e_j in S are related by r_k. The function $f(\cdot)$ itself can be built using any discriminative model such as perceptron or support vector machines (SVMs). The input to this function is a set of features extracted from the sentence. Syntactic and semantic features extracted from positive and negative examples in the training data are used to train the model. These features often consist of the entities themselves, the sequence of words between the entities, the semantic type of the entities, the path connecting the entities in the parse tree, etc. In such approaches, only some of the explored features may be discriminative. It is important to select only the "good" features. The feature-selection process, however, is largely heuristic, and obtaining an optimal set of features is often extremely difficult. To overcome this problem, two lines of work have emerged in relation extraction: one exploring specialized kernels capable of directly exploiting rich structures like parse trees, and the other utilizing continuous space word representations.

The kernel-based methods for relation extraction have been based on string kernels used for text classification by Lodhi et al. (2002). For two data points x and y, the kernel similarity function $K(x, y)$ is used by SVMs (or other discriminative classifiers). Thus, given a training sentence $S = w_0 \dots e_i \dots e_j \dots w_{n_s}$ and a test sentence $T = w_0 \dots e_i \dots e_j \dots w_{n_t}$, the classifier performance effectively depends on how the similarity between S and T is computed. Mooney and Bunescu (2005), for instance, divide their sentences into three portions labeled *before*, *middle* and *after*, and compute three separate kernels for these portions. The final kernel function is simply a sum of these three measures.

Zelenko, Aone, and Richardella (2003) modified the original string kernel function to compute structural similarity between constituency parse trees. In their work, the computations were recursively carried out on subtrees. Culotta and Sorensen (2004) used a very similar kernel to perform the relation extraction task using dependency parse trees instead. They argued that it was the use of richer syntactic and semantic structures that yielded significant performance gains. In contrast, Bunescu and Mooney (2005) observed that using the shortest path connecting two entities in a dependency parse tree is sufficient for relation extraction. Their shortest path kernel, in addition to simplifying the kernel function computation, improved recall while obtaining precision comparable to Culotta's tree kernel.

With the revived interest in neural networks, using continuous space representations has been the other approach intended to overcome the dependency on feature design. These methods rely on learning distributed representations of words, called *word embeddings* (Turian, Ratinov, and Bengio, 2010). Going beyond words, Socher et al. (2012) used a recursive neural network (RNN) to learn embeddings for syntactic tree paths connecting the target entity-pair. Hashimoto et al. (2013), too, used RNN for supervised relation classification. Their work shows that if important cue phrases can be explicitly weighted, we can achieve significant improvements. Subsequently, Zeng et al. (2014) and Zeng et al. (2015) used a convolutional neural network (CNN) with multiple layers of embeddings for words and sentences to improve upon the RNN results. Further, a recent work by Santos, Xiang, and Zhou (2015) showed that a single layer of word embeddings achieves state-of-the-art results when CNN is modified to use a pairwise ranking method. In contrast, RNN and Zeng's CNN models perform multi-class classification using a softmax function.

Supervised methods, however, have one obvious drawback: they require gold-standard labeled data. This is inherently expensive, more so in a specialized domain like scientific research literature where domain expertise is a prerequisite for annotation. As a result, semi-supervised methods are becoming increasingly popular. This body of work is discussed next.

Input: A set of unlabeled data D and a set of seed examples S
repeat
  train a classifier C on S;
  label D using C;
  $N =$ top n labels provided by C (ranked by confidence);
  $S = S \cup N$;
  $D = D \setminus N$;
until *convergence criteria reached*;

Algorithm 1.1: The general structure of Yarowsky's bootstrapping algorithm (Bach and Badaskar, 2007).

Semi-supervised Methods

Lack of sufficient gold-standard labeled data is an obvious impediment to supervised relation extraction. Early notable attempts to circumvent this problem include semi-supervised methods such as DIPRE (Dual Iterative Pattern Relation Expansion) (Brin, 1999) and Snowball (Agichtein and Gravano, 2000). These two methods, and semi-supervised methods in general, have much in common. They are closely related to an earlier algorithm proposed by Yarowsky (1995), presented in Algorithm 1.1. The main idea is to *bootstrap*, *i.e.*, iteratively expand the set of seed relations while taking care to limit the "semantic drift"[4].

DIPRE represents every seed element as a triple of contexts: *before*, *between*, and *after*. It then generates extraction patterns by using string matching to group these contexts. The semantic drift is controlled by limiting the number of instances each pattern can extract. Snowball, too, uses these three contexts, but uses a TF-IDF vector representation of these contexts followed by a one-pass clustering based on cosine similarity. The cluster centroids are treated as valid extraction patterns. By not requiring exact surface matching like DIPRE, Snowball allows for some flexibility. The patterns are scored and ranked, and only the instances that pass a threshold are used as seed for the next iteration. Continuing to explore more robust ways of controlling semantic drift while learning general patterns, others have employed word clusters (Sun, Grishman, and Sekine, 2011) and in the very recent BREDS system (Batista, Martins, and Silva, 2015), word embeddings.

The approaches discussed above all use clustering algorithms to select the instances that get included in subsequent iterations. Some approaches have also used bootstrapping with SVMs (*e.g.* Zhang (2004)), thus formulating a semi-supervised technique closer in spirit to the supervised approaches discussed earlier. Another body of work has exploited label propagation algorithms instead. Fundamentally, the manifold structure (*i.e.*, the cluster structure) of the underlying data is still used, but is exploited by defining a graph where (possibly weighted) edges reflect similarity, rather than resorting to an explicit clustering algorithm. This can be seen in the works of Chen et al. (2006) and GuoDong, LongHua, and QiaoMing (2009), among others.

Distant Supervision

As a relatively recent approach to alleviating the problem of lack of labeled data, distant supervision has become increasingly attractive. It has been used with success in other domains by Mintz et al. (2009) and Hoffmann et al. (2011), among others, often using Freebase[5] as the knowledge

[4] Semantic drift is the evolution of the usage of a word or phrase that results in the meaning of the word itself to change (McIntosh and Curran, 2009)
[5] Bollacker et al., 2008.

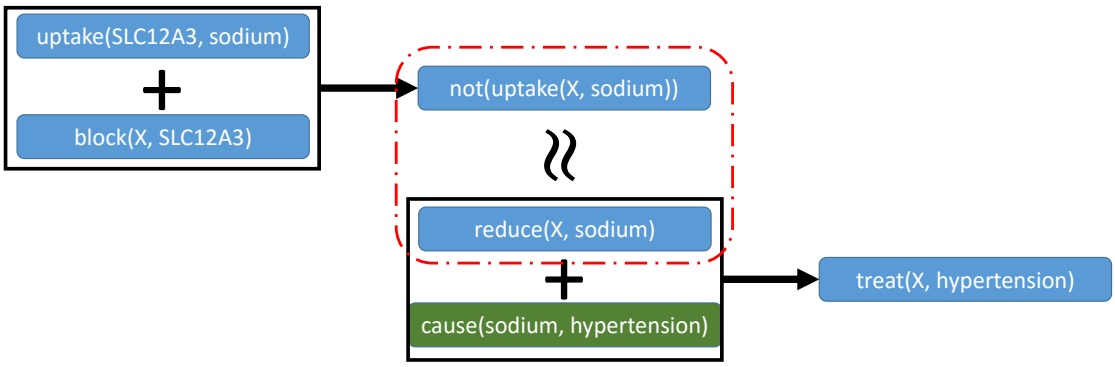

Figure 1.2: A simple two-step relation inference process illustrating how an unknown drug X may be discovered to be a potential treatment for hypertension based on (i) an *intra-document* inference across multiple sentences leading to the relation $\neg(\texttt{uptake}(X, sodium))$, (ii) semantic similarity of predicate-argument structures, and (iii) an *inter-document* inference that combines with the prior knowledge $\texttt{cause}(sodium, hypertension)$. This particular illustration is based on the *pathway* of hydrochlorothiazide, which treats high blood pressure by blocking the action of the SLC12A3 gene.

source. In recent years, the biomedical domain has also seen the use of distant supervision using databases such as the Yeast Protein Database[6] (YPD) to detect relations involving proteins (Craven and Kumlien, 1999), IntAct[7] to extract protein interactions (Thomas et al., 2011) and the Unified Medical Language System[8] (UMLS) for various medical relations (Roller and Stevenson, 2014).

Distant supervision, however, provides data that is inherently noisy. Prior research, *e.g.* Mintz et al. (2009), has often worked with the assumption that if two entities e_i and e_j are known to be in a relation r_k, then any sentence containing both entities express r_k. As described in Sec. 1.1, this assumption is too strong for our domain. Previous methods (*e.g.*, Mintz et al. (2009), Riedel, Yao, and McCallum (2010), Hoffmann et al. (2011), Surdeanu et al. (2012)) have typically applied the supervised learning paradigm with feature designs to obtain distantly supervised labeled data. To reduce the noise in such data, these methods model relation extraction as a *multi-instance* learning problem: the assumption that if multiple sentences contain e_i and e_j, the relation r_k holds in at least one of them. The features are aggregated from multiple sentences where the target entity-pair is present. The aggregation is accomplished by the use of latent variables to represent sentence-level decisions.

A major limitation of all the relation extraction approaches is that they extract relations at the sentence level. Relations can, in fact, span over multiple sentences and even across multiple documents. Such relations are better *inferred* than directly extracted. Much of prior research has modeled this inference process as learning from knowledge graphs, which we discuss next.

1.2.2 Relation Learning from Knowledge Graphs

In natural language texts, it is often the case that the cues of a relation between two entities are not contained within a single sentence. Particularly so in biomedical research literature, as described earlier in Sec. 1.1. These relations may, however, be *inferred* based on various criteria if the problem is modeled as a graph. In statistical relation learning, such graphs that represent knowledge bases

[6] Hodges, Payne, and Garrels, 1998.
[7] Hermjakob et al., 2004.
[8] Lindberg, Humphreys, and McCray, 1993.

(KBs) are called knowledge graphs (KGs). YAGO[9], NELL[10] and Freebase are some of the well-studied examples. In this section we will look at how relation learning, which pertains to creating models for relational data, can be applied to KGs. Since our focus remains on binary relations r_k between two entities e_i and e_j, relations will be denoted by triples of the form (e_i, r_k, e_j). This is consistent with the RDF standard of relation representation using $(subject, predicate, object)$ (SPO) triples. Given a KG, one may assume one of two positions: triples absent from the graph (a) indicate false relations, and (b) are treated as unknown facts. These are known as the *closed-world* and *open-world* assumptions, respectively. Most KGs are known to be highly incomplete, especially so in the biomedical domain due to the rapid growth in data. Our work, therefore, embraces the open-world assumption.

Knowledge graphs typically obey certain hard constraints like type contraints, transitivity, etc. In the biomedical domain, UMLS may be viewed as a KG, albeit sparse due to relational information being highly incomplete. Additionally, such graphs also often exhibit *soft constraints*, some of which are quite important with respect to our work:

(1) **Homophily, or *autocorrelation*** states that any entity exhibits a tendency to be related to other entities with similar characteristics (Nickel, 2013). Homophily is known to be present in many relational datasets (Jensen and Neville, 2002), and can be a powerful predictor of unknown relations. For example, we can predict a new adverse effect of a drug by studying the adverse effects of the drugs similar to it.

(2) **Long-range dependency** over paths in a graph of known relations can yield new relations. These paths can be expressed as logical conjunctions of the form $(e_1, r_1, e_2) \wedge (e_2, r_2, e_3) \ldots \wedge (e_{k-1}, r_{k-1}, e_k) \implies (e_1, r_k, e_k)$. Fig. 1.2 shows a simple 2-hop inference for a drug that may treat hypertension. The corresponding logical formalism is $(x, \text{reduce}, \text{sodium}) \wedge (\text{sodium}, \text{cause}, \text{hypertension}) \implies (x, \text{treat}, \text{hypertension})$.

Needless to say, a biomedical KB must heed two extremely important parameters that govern its usefulness – *completeness*, and *accuracy*. As observed by Nickel et al. (2015), constructing such a KB may involve methods that fall under one of four categories:

- manual curation, which is how medical KBs are currently created and expanded,
- collaboration, where relation triples are created manually by an open group of volunteers,
- automated semi-structured methods, where triples are extracted from semi-structured texts using rules and regular expressions, and
- automated unstructured methods, extracting triples from texts using NLP techniques.

Note that given the specialized nature of our domain, open collaboration is not a feasible option. Also, while later chapters of this thesis adopt automated semi-structured techniques, such techniques require semi-structured data to already exist. In the early stages of the translational pipeline, KB construction and expansion must exploit unstructured data as the primary source of new knowledge. From this perspective, our work on relation inference may be viewed as a link prediction task – *i.e.*, to predict whether or not an edge exists – aimed at expanding a KB.

In the remainder of this section, we will denote each *possible* relation as the triple $x_{ijk} = (e_i, r_k, e_j)$ over the set $\mathcal{E} = \{e_1, \ldots, e_n\}$ of entities and the set $\mathcal{R} = \{r_1, \ldots, r_m\}$ of relations, where $|\mathcal{E}| = n$ and $|\mathcal{R}| = m$. The triples are then modeled as binary random variables

$$y_{ijk} = \begin{cases} 1 & \text{if } x_{ijk} \text{ exists} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

[9] Suchanek, Kasneci, and Weikum, 2007.
[10] Carlson et al., 2010.

Relation Learning Models

The presence (absence) of some relations often indicate the presence (absence) of certain other relations, indicating that global inference models are particularly well suited for relation learning in knowledge graphs. The inter-dependence between various relations, *i.e.*, the correlations among the y_{ijk} variables, can be modeled in many ways. In the knowledge graph paradigm, these can be thought of as models exploiting (i) *latent features*, and (ii) *graph features*. These two categories assume conditional independence of the y_{ijk} variables given latent and observed features, respectively. Before describing some particular models relevant to our work, we present a high-level formalism as follows:

$$P\left(y_{ijk}|\mathcal{R}_{known}, \Theta\right) = B\left(y_{ijk}|\sigma\left(f\left(x_{ijk}; \Theta\right)\right)\right) \tag{1.2}$$

where $\mathcal{R}_{known}$ is the set of known relations, Θ is the set of model parameters, f is a scoring function representing the model's confidence that the relation x_{ijk} is true, $\sigma(x) = 1/\left(1 + e^{-x}\right)$ is the sigmoid function, and $B(\cdot\,|\,\cdot)$ denotes the Bernoulli distribution. Equation 1.2 provides the general structure of probabilistic models, but other score-based models can be transformed into this form by Platt scaling (Platt, 1999; Niculescu-Mizil and Caruana, 2005).

Latent Feature Models

In this class of models, the *entities* are modeled by latent variables. The probability of a relation between two entities is then computed based on operations on these latent variables, as shown by Kok and Domingos (2007). Others have explored assigning a latent *class* to each entity (Kemp et al., 2006; Xu et al., 2012). This means that the latent variables are binary and mutually exclusive (since each entity can have exactly one class label). This body of work is based on the idea that entities will first be assigned classes, and the probability of a relation between e_i and e_j will then be derived from the probability of the relation existing between their respective latent classes. Subsequently, Airoldi et al. (2006) generalized this method to allow entities to obtain multiple labels, thereby abandoning the requirement for the latent variables to be mutually exclusive.

Recently, bilinear vector space models have been shown to outperform significantly richer parametrized models in relation learning. For example, RESCAL (Nickel, Tresp, and Kriegel, 2011) achieved state-of-the-art performance on several datasets, and a modified bilinear model, DISTMULT (Yang et al., 2015), has achieved better results than richer models like TRANSE (Bordes et al., 2013). In this approach, relations are modeled via pairwise interaction of latent variables, with the score function

$$f_{ijk} = \mathbf{e_i}^{\mathrm{T}} \mathbf{W_k} \mathbf{e_j} \tag{1.3}$$

where $\mathbf{W_k}$ is a weight matrix representing the k^{th} relation r_k, and $\mathbf{e_i}$ and $\mathbf{e_j}$ denote the latent feature representations of the corresponding entities e_i and e_j. The magnitude of the diagonal entries of $\mathbf{W_k}$ indicate the degree of homophily (Nickel, 2013).

Graph Feature Models

Now, we present a brief discussion of models that learn relations from the actual edges in the knowledge graph itself, without resorting to hidden variables. From this body of research, our relation inference is most similar in spirit to the cross-document information integration method of Yao, Riedel, and McCallum (2010) using factor graphs, the inference based on a path ranking algorithm (PRA) by Lao, Mitchell, and Cohen (2011) and the work on relation entailment graphs by Berant, Dagan, and Goldberger (2012).

While these methods provide powerful relation entailment capabilities, they typically infer more relations for an unchanging entity-pair. For instance, given two drugs d_1 and d_2, we could obtain an entailment like $(d_1, \texttt{derive}, d_2) \implies (d_1, \texttt{process-from}, d_2)$. In this domain, however, our focus is on inferring a relation between an entity-pair based on *other* related entities. This is unlike the above mentioned body of work where the focus is to *drill down* from general to more specific relations between the *same* entities.

Very recently, Gardner and Mitchell (2015) have built on PRA to show that entity-neighborhoods in graphs can be used for knowledge based completion. Their technique, called *subgraph feature extraction*, is similar to PRA in the sense that they too generate feature matrices over node pairs in the knowledge graph. Each column of such a matrix is a path type in the graph, and each row corresponds to a node-pair. Another recent work suggesting a PRA-like model is the compositional vector space model proposed by Neelakantan, Roth, and McCallum (2015). Their work showed that path bigram features (*i.e.*, paths of length 2 in the graph) resulted in significant improvement.

Our work, too, models such paths in knowledge graphs. The major distinction, however, is that in our problem, the intermediate edges are seldom present in the knowledge graph. Therefore, they must be modeled implicitly. Our first model, presented in Sec. 1.5.1 is similar to a latent modeling of the path bigram features explored by Neelakantan, Roth, and McCallum (2015). We then generalize this approach to model long-range dependencies in an integer linear programming framework (presented in Sec. 1.5.2).

1.3 Motivation

The key observation that motivates our approach is that research articles describing drugs, diseases and their relationships, often do so in terms of the underlying *physiological* effects. As noted before, in many cases, the effects of a drug X do not bear any explicit linguistic connection to a disease Y. As a result, a relation $r(X, Y)$ must be *inferred* based on two things:

(a) the physiological effects $E_X = (e_1, e_2, \ldots, e_n)$ of X, and
(b) the characteristics $C_Y = (c_1, c_2, \ldots, c_m)$ of the disease Y.

The inference process itself can be thought of as an implicature where X is deemed beneficial if it provides evidence of opposing the characteristics of Y, *i.e.* $e_i = \neg(c_j)$. Similarly, X should be considered harmful if it aids the disease characteristics. The above example is an overly simplistic depiction. In most cases, we find that there are several layers of physiological effects that need to be semantically linked before establishing a drug-disease relation. Conceptually, this outlook is similar to the use of implicature in the graph-based sentiment propagation method proposed by Deng and Wiebe (2014). Figure 1.2 illustrates this concept by showing how a drug-gene and a gene-chemical relation combine with prior knowledge to infer a potential therapeutic relation.

While current state-of-the-art relation learning methods are capable of long-range inferences in a knowledge graph, they still require the intermediate steps to be a represented in the graph. However, we note that often these intermediate steps that express the physiological effects, are missing from the graph. Thus, there is a need to incorporate them as latent variables in order to infer how entities such as drugs and diseases are related. This raises the first major distinction between our work and prior research using latent variable models for relation learning. Instead of building latent feature representations of the entities, our latent features *are* entities and relations. In particular, they are entities and relations extracted from a text corpus, and not necessarily a part of the graph.

Further, since these variables are extracted from unstructured text using standard NLP tools, they are subject to erroneous extraction. As a result, directly adding this extracted knowledge yields an extremely noisy graph, and any subsequent relation learning suffers. To alleviate this, we adopt two approaches to global inference.

The first is a *one-hop* maximum likelihood inference. If we were to add all the extracted entities and relations to the original KG, this method could be viewed as relation learning restricted to paths of length 2 in that graph. The second is an integer linear programming (ILP) framework that exploits compatibility constraints to impede noisy propagation. The constraints we encode can be likened to the manifold regularization constraints used by Wang and Fan (2014) in their sentence-level relation extraction system for medical relation extraction.

In biomedicine, the observation that a drug and a disease could be therapeutically related because of their opposing physiological effects, was first made by Swanson (1986). His initial discovery was that fish oil (X) may treat Raynaud's disease (Y) because the latter is characterized by high platelet concentration and high viscosity in blood, while the former reduces both. In that work, he points out that this relation was thus far unknown because of "noninteracting literatures" in the sub-domains within medical researchers. This was a completely manual process. In the following years, some text mining systems were developed to aid this kind of knowledge discovery (*e.g.* Arrowsmith[11], BITOLA[12] and LitLinker[13]). These, however, simply use co-occurrence as a potential indication of a relation – the actual existence of a relation, and what that relation might be – is left to the domain-experts (Hristovski, Rindflesch, and Peterlin, 2013). More recent research has used rule-based approaches to incorporate the *semantics* of a relation. Notable examples include the SemRep[14] program built by Rindflesch and Fiszman (2003) and a combination of rules and co-occurrence measures proposed by Hristovski et al. (2006). These, too, are largely manual. Further, their inference procedure is *local* in the sense that the rules do not incorporate the chain of implicit effects beyond a single step. To summarize, the limits of these knowledge discovery methods are that

(a) significant manual involvement is required,
(b) they are restricted to a *one-hop* inference, and
(c) the knowledge discovery for intermediate physiologic effects remains at the clause-level.

To overcome these shortcomings of current methods, and to stay true to the relation inference paradigm described so far, we delve deeper into some terminology borrowed from the biomedical sciences. These definitions provide the basis of our relation inference method formally described in Sec. 1.4.

1.3.1 A pharmacologic perspective

In most biomedical research articles, the natural language narrative is an explanation of not just the *what*, but also the *why* and *how* of observed phenomena. The explanation of events in terms of physiological effects, thus, is to be expected. Biomedical researchers as well as healthcare practitioners have a keen interest in understanding the mechanism of the action of a drug – be it an adverse effect or a therapeutic one. Even for non-pharmacological entities (micro-organisms, chemicals, etc.), due to scientific interest as well as strong economic considerations, this understanding is of fundamental importance in biomedicine.

[11] Smalheiser and Swanson, 1998.
[12] http://ibmi3.mf.uni-lj.si/bitola/
[13] Pratt and Yetisgen-Yildiz, 2003.
[14] http://semrep.nlm.nih.gov/

Table 1.1: The two "super"-relations and their constituent fine-grained UMLS types.

`beneficial (B)`	may_treat, may_prevent, treats, prevents
`harmful (H)`	cause_of, causative_agent_of, contraindicated_drug

The detailed effects described in biomedical research literature broadly fall under two categories, each a discipline in its own right:

(a) *pharmacokinetics*, the study of drug absorption, distribution, metabolism, and excretion (Ratain and Plunkett, 2003), and

(b) *pharmacodynamics*, the study of the biochemical and physiological effects of drugs and their mechanisms of action (Brunton, Lazo, and Parker, 2005, ch. 3).

In simpler terms, the former is the study of what happens to the drug in the body, while the latter is the study of what the drug does to the body.

Drug action is the end result of several branched or consecutive reaction steps (Seydel and Schaper, 1981). The sequence of these steps is called the *pathway*. From a purely pharmacologic perspective, the sequence of intermediate entities and effects that we exploit in order to infer the final relations between two entities are usually pharmacodynamic/pharmacokinetic pathways. In other words, the methodology we describe next, works by a latent modeling of these pathways. A potential advantage of this is that a slight modification of our core inference framework allows for a characterization where the steps of these pathways can be made more explicit. Even though that makes the inference process computationally more expensive, it provides an inference model that can better explain *why* two entities share a particular semantic relation.

In the remainder of this chapter, we present our methodology in Sec. 1.4, followed by the formal definition of our inference models in Sec. 1.5. Next, Sec. 1.6 discusses the experimental setup, along with the results we have obtained so far. The chapter then concludes with the evaluation and error analysis in Sec. 1.7.

1.4 Methodology

As discussed previously in Sec. 1.1, standard sentence-based extraction approaches have limited efficacy in this domain. The sentences tend to be long and complex, making it hard to learn generalizable lexico-syntactic extraction patterns. Further, a significant proportion of relations are not directly expressed within a single sentence. Thus, instead of extracting relations, we focus on an *inference* paradigm wherein pieces of relevant information expressed in multiple sentences and/or documents are combined. Our approach identifies a relation between two entities by modeling how they are related to other relevant entities. Specifically, we observe that many narratives mention the effects without directly stating the final outcome. For example, most drugs that treat diabetes often target a reduction in blood glucose levels even though they may differ in the mechanisms and sites used to achieve this effect. From a medical standpoint (see Sec. 1.3.1), this can be viewed as an aggregation of the pharmacodynamic/pharmacokinetic pathways of these entities. Formally, the entailment may be expressed as $\texttt{lower}(x, glucose) \Rightarrow \texttt{treat}(x, diabetes)$. However, since there are no readily available knowledge bases (KBs) that provide this kind of gold-standard relation entailment data, we develop a framework that models the pathways *implicitly*[15].

We use two basic ideas to discover relations. First, we find the most frequent subject-verb-object (SVO) triples where a drug is the subject, and we treat the VO pairs as implicit effect in the drug's pathway. To reduce noise, we use domain knowledge to induce semantic type constraints on the

[15] These pathways are implicit, or latent, because they are neither observed in the model nor available during training.

objects. For example, if we know the relation $\texttt{treats}(insulin, diabetes)$ and we extract the SVO $\texttt{lower}(insuling, glucose)$, the VO pair $\texttt{lower}(glucose)$ is treated as a latent effect. This allows us to infer that any new drug that lowers glucose is also likely to have a therapeutic effect on diabetes, even when there is no direct sentence-level evidence that explicitly conveys such a relation.

Second, we use a global inference model to discover new classes of drugs whose targets may have been unobserved in the training data. From a pharmacologic perspective, this means that our model is capable of identifying new pathways that were not available in the training data. This model is formulated as in integer linear programming (ILP) problem, which encourages new drugs that achieve similar pharmacodynamic effects as known drugs, to be assigned compatible relations.

Given a disease, our goal is to identify (i) new therapeutic options for it, and (ii) entities that may be harmful for patients afflicted with it. By therapeutic options, we mean drugs or procedures that may be used to treat, prevent or manage a disease or its symptoms. Similarly, we also want to identify drugs, procedures or even lifestyle aspects (*e.g.* smoking) that may cause or exacerbate a disease or its symptoms. Such an inference system can aid the translational pipeline by allowing quick augmentation of resources like UMLS or DrugBank[16] that provide critical knowledge about biomedical concepts. To this end, we follow the distant supervision approach: use UMLS as the knowledge source, and assume that we are not given any sentence-level annotations.

Note that biomedical KBs like UMLS provide fine-grained relation labels. Since our goal is to identify therapeutic and adverse-effect relations, we follow the approach adopted by Wang and Fan (2014), and coalesce some of these fine-grained relations into two "super"-relation categories: `beneficial` and `harmful`. Table 1.1 lays out the correspondence between these super-relations and their fine-grained UMLS constituents. The problem definition can be stated thus

> Given some prior knowledge in the form of a set of diseases $D = \{d_1, \ldots, d_n\}$ and a set of drugs $R = \{r_1, \ldots, r_m\}$ such that $\forall\, i \in [1, m], \exists\, j \in [1, n]$ for which either $\texttt{beneficial}(r_i, d_j)$ or $\texttt{harmful}(r_i, d_j)$ holds, discover new drugs – by means of inferences drawn from biomedical research literature – that bear either a `beneficial` or a `harmful` relation to a disease in D.

1.4.1 The *latent pathway* model

Since we only have drug-disease relations as prior knowledge, we model the pharmacodynamic/ pharmacokinetic pathway of a drug as latent knowledge. For this, an initial analysis was carried out on sentences with drug mentions. Due to the complexity of the sentences, we extracted predicates that were in the dependency neighborhood of the drugs and the objects that they modify. These SVO triples constitute individual steps in the drug's pathway, and were treated as *candidates*. We found that assessing these candidate steps as beneficial, harmful or unrelated to a disease was extremely challenging as it required domain-specific semantic knowledge of various entities. Returning to the example presented in Fig. 1.2, $\texttt{block}(hydrochlorothiazide, SLC12A3)$ is such a candidate. Understanding it as a step that is beneficial for the target disease, hypertension, requires the knowledge that it is a gene that is responsible for reabsorbing sodium into the body.

We choose a simpler approach where we do not explicitly classify the effects as beneficial or harmful. Instead, we gather the most frequent predicates and their objects as a surrogate for the key steps in a drug's pathway. This allows us to convert the extraction problem into a similarity-based inference problem, where drug-drug similarity is measured in terms of their prominent pharmacological effects. Next we describe the process of extracting these steps, and present two

[16] Knox et al., 2011.

Treatment	antibiotic, clinical drug, hazardous or poisonous substance, organic chemical, pharmacological substance, steroid, vitamin
Disease	acquired abnormality, anatomical abnormality, congenital abnormality, disease or syndrome, cell or molecular dysfunction, neoplastic process, pathologic function, sign or symptom
Theme	anatomical structure, body location or region, body part, organ or organ component, body space or junction, cell component, cell, laboratory or test result, biologic function, cell function, genetic function, molecular function, organism function, organ or tissue function, physiologic function, amino acid, peptide or protein, enzyme, hormone

Table 1.2: Coarse semantic categories and their constituent fine-grained UMLS semantic types.

global inference formulations that use them to identify new drugs pertaining (either as harmful or as potentially therapeutic treatments) to a given disease.

Extracting the "steps" of a pathway

We use the PubMed Central (PMC) repository for relation extraction from biomedical research literature. To extract mentions of drugs, diseases and other biological or physiological entities, we use Genia (Tsuruoka et al., 2005), a tagger and parser for biomedical texts, to obtain phrase chunks of the dataset. Subsequently, each noun phrase was provided as input to a well known medical named entity recognizer, MetaMap (Aronson, 2001), which uses the UMLS resource to identify each phrase to one or more entities. Additionally, it labels each identification with a *semantic type*, obtained from UMLS. The UMLS Metathesaurus consists of 133 such semantic types, and uses these types to categorize more than 2 million entities[17]. We use this fine-grained semantic type information to determine whether or not a phrase mentions a relevant entity.

Entities relevant to our approach are primarily drug and disease mentions. But we also want to extract any biological and/or physiological entities that may either affect or be affected by a drug or a disease. In this proposal, we call such an entity a *pharmacologic theme*, shortened to just "theme". We collect a small subset of the fine-grained UMLS categories to define three coarse semantic types: *treatment*, *disease* and *theme*. These coarse semantic types and their constituent UMLS categories are shown in Table 1.2.

MetaMap often maps a phrase to multiple entities, assigning a score to each mapping. In our work, we experimentally decided upon a cutoff value, and considered only those candidates that had score beyond this threshold. We also observed that biomedical literature contains an extremely high number of context-dependent non-standard abbreviations. To correctly resolve them, we identified their first mention in a document, and applied a well-known abbreviation resolution algorithm for this domain (Schwartz and Hearst, 2003) to obtain the correct expansion, and subsequently, the correct semantic type as determined by MetaMap. In this manner, each document was labeled with sentence- and phrase-level entity mentions. In case of a conflict between the coarse-types of an entity mention (*e.g.* "Insulin" is a theme as well as a treatment), we chose to record both types.

[17] *UMLS® Reference Manual [Internet]* (2009, Ch. 5). Available: http://www.ncbi.nlm.nih.gov/books/NBK9679

It should be noted that a vast majority of medical entities have a variety of synonymous names, as noted earlier in Sec. 1.1. To normalize such surface differences, we also recorded the canonical name (provided by the UMLS Metathesaurus) for each entity. For example, both "shortness of breath" and "dyspnea" are identified with the latter, which is its canonical name in UMLS. After identifying the relevant medical entities, we extracted triples of the form `predicate(treatment, theme)` using dependency parse trees. All the sentences were parsed using ClearParser (Choi and Palmer, 2011), which provides a model trained on biomedical language. Then, the dependency paths connecting a drug-disease pair or a drug-theme pair were extracted. These paths in turn were used to identify and extract predicates through which a drug may affect a theme.

In the predicate extraction step, we considered not just verbs, but also nominal predicates, *i.e.* verbs that are used in their noun forms. This is a common usage pattern, evident in clauses of the type "insulin may have caused a reduction in serum glucose". In such cases, we *de*-nominalized the predicates (*e.g.* "reduction") that satisfied *each* of the following conditions:

(a) the nominal predicate does not have a noun compound modifier
(b) is itself not a noun compound modifier of another noun
(c) has a preposition child node in the dependency path

1.5 Global Inference

As discussed in the related work (see Sec. 1.2), knowledge discovery from biomedical research literature must deal with a scenario where the complete knowledge may be fragmented across multiple sentences in a single document, or even across multiple documents. The relation inference process must therefore be able to incorporate *global* cues. Further, such cues must be combined correctly based on the similarity of drug pathways in order to obtain the final relation. In this section, we first present a simple *one-hop* maximum likelihood inference (MLI) method that stays true to this goal. We then describe a more sophisticated modeling of global inference using an integer linear programming (ILP) framework.

1.5.1 Maximum Likelihood Inference

Here we describe a simple score-based method for relation inference. Our objective is to score new drugs based on their likelihood of being beneficial (harmful) for a given disease, based on the overlap of their effects with the effects of the known beneficial (harmful) drugs. Formally, we represent a *pharmacologic effect* of a drug by an (`action`, `theme`) pair $e = (p, t)$, where p denotes a pharmacologic action on a theme t. The pair (`reduce`, `blood glucose`) is an example of such an effect[18]. We will denote the set of drugs, diseases and pharmacologic themes by $R = \{r_1, r_2, \ldots, r_m\}$, $D = \{d_1, d_2, \ldots, d_n\}$, and $T = \{t_1, t_2, \ldots, t_k\}$, respectively. Our goal is to infer "super"-relations (see Table 1.1) $s(r, d)$ between drugs and diseases, where $s \in S$, and $S = \{\text{B}, \text{H}\}$. Further, let $R_{d^+} \subseteq R$ and $R_{d^-} \subseteq R$ denote the sets of drugs known (from prior knowledge) to be in B and H super-relations, respectively, with respect to a disease $d \in D$.

For every drug r in prior knowledge, we extract all mentions of an effect e of r and compute the probability[19] of observing e within all drugs exhibiting the same super-relation (`beneficial`

[18] Note that what we are defining here is, from a medical perspective, a single step in the drug's pharmacodynamic pathway.

[19] Strictly speaking, we are computing the relative frequency. But since we can convert this into a probability measure via Platt scaling (Platt, 1999), we are using the term *probability* for ease of expression.

or `harmful`) toward a disease $d \in D$. Denoting the frequency of a drug r being the subject of an effect e by $\nu(r, e)$, for the beneficial super-relation, this can be computed as

$$Pr(e|R_{d^+}) = \frac{\Sigma_{r \in R_{d^+}} \nu(r, e)}{\Sigma_{r \in R_{d^+} \cup R_{d^-}} \nu(r, e)} \tag{1.4}$$

To discover new relations between drugs and the disease d, we first find all drugs and their effect mentions by looking for (r, p, t) triples where (p, t) denotes the effect as before. Then, we compute the likelihood of r_i being beneficial for d as

$$\forall r_i \notin R_{d^+} \ , \quad \mathcal{L}_{r_i \in R_{d^+}} = \sum_j \left(Pr\left(e_j | R_{d^+}\right) . \nu(r_i, e_j) \right) . \tag{1.5}$$

Interchanging R_{d^+} and R_{d^-} in equations 1.4 and 1.5 yields the corresponding formulas for inferring harmful super-relations.

In spite of its simplicity, this formulation achieves two key goals. First, it attends to the syntactic complexity of biomedical sentences by reducing the problem to the space of predicates and themes. Second, it goes beyond sentence-level relation extraction by aggregating information across different mentions of the effects and drugs. These mentions may be across multiple sentences in the same document, or even across multiple documents in the dataset.

Extensions of the maximum likelihood inference approach

Even though the above approach is able to learn from a global distribution of features, it is restricted in the sense that only those drugs that have some effects identical to the drugs in prior knowledge can be discovered by it. Therefore, in the event that a drug treats a disease, but does so in a way that has little in common with the drugs in prior knowledge, MLI in the last section will fail to detect the new drug. Further, it suffers from a limitation on the *number of inference steps* – much like the manual literature-based discovery method of Swanson (1986), the simple likelihood estimation can only do a *one-hop* inference. If the drug pathways are never stated in a manner so simple that there is only one intermediate entity affected by the drugs, the previous approach will not be able to discover new relational information. There are, however, several simple ways to augment this approach. The focus of this thesis lies in the translation pipeline as a whole, and as a result, a detailed discussion of such extensions is beyond the scope of this work. But, before presenting our second approach using ILP, we will digress briefly to sketch the outlines of two other possible directions.

The first is to place it in context of Yarowsky's bootstrapping algorithm (see Algorithm 1.1). Given the seed set of beneficial and harmful drugs, MLI can be used instead to rank the other drugs. Subsequently, the top few can be chosen based on a threshold value, and added to the seed set for the next iteration. This approach, however, is likely to suffer from poor performance unless carefully designed constraints and convergence criteria are added. Experimental results from prior research has shown that iterative self-training has a tendency to quickly succumb to semantic drift (Komachi et al., 2008; McIntosh and Curran, 2009). But designing good constraints on a simple bootstrapping approach is a largely heuristic process.

Perhaps a more natural way to extend this kind of a *one-hop* process is view it as a perceptron, and subsequently design a feed-forward neural network with multiple layers instead of just one. Constraints can be imposed naturally via excitation and/inhibition of neurons based on various criteria. A simple illustration of this is shown in Figure 1.3. Of particular interest here is that we

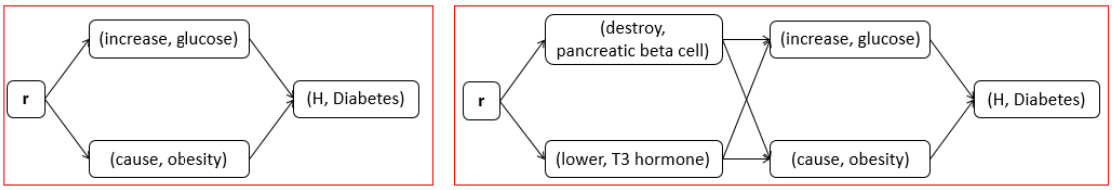

Figure 1.3: The one-hop MLI process (left) extended (right) to a multi-layer feed-forward neural network: a chain of effects can then propagate through the layers, ultimately determining whether or not the drug r has a harmful super-relation (shown here with H) with diabetes.

may not need to *learn* the weights because if we train a neural network using mean square error or cross-entropy cost function, the outputs will converge to the posterior class probabilities (Gish, 1990; Richard and Lippmann, 1991). But these values can be induced directly from co-occurrence statistics in the corpus.

1.5.2 A formulation based on Integer Linear Programming

In this work, we chose to address the shortcomings of the MLI approach by designing a global inference algorithm based on integer linear programming (ILP). In contrast to the one-step likelihood estimation method, this is a setting that imposes inter-dependent constraints on drugs based on how (dis-)similar their effects are to those of other drugs. To strengthen the constraints, we include two new components: (i) prior knowledge of drugs in opposing super-relations, and (ii) lexical similarity measures for distinct predicates.

Distance Measurements

For each drug $r_i \in R$, we collect all its effects (r_i, p_j, t_k), *i.e.* all triples of the form `lower(insulin, glucose)`, and measure its likelihood by computing the relative frequency measure

$$f_{i,j,k} = \frac{n_{i,j,k}}{N}, \tag{1.6}$$

where $n_{i,j,k}$ denotes the frequency of the triple (r_i, p_j, t_k), and N is the total count of all such triples representing drug-effect mentions. We impose a notion of drug-drug similarity based on the *similarity of the probability distributions of their effects*. This measure, however, cannot be directly imposed because it fails to take into account the lexical (dis-)similarity of the predicates that appear in these effects. For example, if one drug *lowers* blood glucose level while another *reduces* it, their effects should be considered as nearly equivalent, even though the predicates are different.

We compute the probability distribution of a drug's effects by constructing different distributions for each theme that the drug acts on. In this way, the difference between two drugs is explicitly modeled in terms of how different their actions on a particular theme are. For instance, if two drugs have similar predicates through which they act on 'blood glucose', but very different predicates for their effect on 'joint pain', they may share a super-relation with respect to one disease (*e.g.* diabetes) but not with respect to another one such as arthritis.

More formally, for two drugs $r_1, r_2 \in R$, let $\mathbb{P}_t(i)$ denote the probability that r_1 acts on t through the predicate p_i, and $\mathbb{Q}_t(i)$ the probability that r_2 acts on t through p_i. If $\mathcal{T}$ is the set of themes on which both drugs r_1 and r_2 exhibit some effect, and $\mathcal{P}_{r_1,t}$ and $\mathcal{P}_{r_2,t}$ are the sets of predicates in those effects, then the difference between the two drugs is expressed in terms of the symmetrized

K-L divergence (Kullback and Leibler, 1951):

$$\Delta(r_1, r_2) = \frac{1}{|2\mathcal{T}|} \sum_{t \in \mathcal{T}} (D_t(r_1, r_2) + D_t(r_2, r_1)) \tag{1.7}$$

where

$$D_t(r_1, r_2) = \sum_{i \in \mathcal{P}_{r_1,t}} \mathbb{P}_t(i) \ln \frac{\mathbb{P}_t(i)}{\mathbb{Q}_t(i)} + \sum_{i \in \mathcal{P}_{r_2,t}} \mathbb{Q}_t(i) \ln \frac{\mathbb{Q}_t(i)}{\mathbb{P}_t(i)} \tag{1.8}$$

The above equation does not account for the lexical similarity of predicates, which we have built into the construction of $\mathbb{P}_t$ and $\mathbb{Q}_t$. To compute the divergence in eq. 1.8, we define the probability distributions in terms of a lexical similarity function σ and the relative frequency. In this work, we employed the WordNet[20]-based similarity score proposed by Wu and Palmer (1994). This is a knowledge-based (as opposed to *corpus-based*) measure of similarity between concepts, rather than words, but it can be easily turned into a lexical similarity metric by selecting, for a given pair of words, those two senses that yield the highest concept similarity score. This, in fact, has largely been the way prior NLP research has employed this similarity metric (*e.g.* McCarthy et al. (2004) and Mihalcea, Corley, and Strapparava (2006)). Denoting by $\delta_f(p_i, p_j, t)$ the difference in the relative frequencies of r_1 causing the effect p_i and r_2 causing the effect p_j on the theme (t), we set $\mathbb{P}_t(i)$ to be the relative frequency (see eq. 1.6) of r_1 acting on t through p_i, and

$$\mathbb{Q}_t(i) = \frac{1}{\mathcal{P}_{r_2,t}} \sum_{p_j \in \mathcal{P}_{r_2,t}} \sigma(i, j) \{1 - \delta_f(p_i, p_j, t)\} \tag{1.9}$$

Note that even though we have defined $\mathbb{P}_t(i)$ and $\mathbb{Q}_t(i)$ differently, the measure of the difference between r_1 and r_2 is symmetric. This is ensured by eq. 1.7.

Incompatibility Constraints

The ILP formulation is designed so that two drugs that, according to eq. 1.7, are similar *enough*, are encouraged to have the same super-relation label. The optimization function associates the cost $\Delta(r_1, r_2)$ to any drug-pair that, for an $\epsilon > 0$, violates the constraint between drug-drug similarity and super-relation labels:

(i) $\Delta(r_1, r_2) < \epsilon$ but r_1 and r_2 have different super-relation labels, **or**

(ii) $\Delta(r_1, r_2) \geq \epsilon$ but r_1 and r_2 have identical super-relation labels.

To formalize this constraint violation, we introduce indicator variables $\delta^{\text{S}}(r_1, d)$ and $\delta^{\text{S}}(r_2, d)$ to denote whether r_1 and r_2 have a beneficial (or harmful) relation with respect to d. Then, a third indicator variable $w(r_1, r_2)$ is defined as

$$w(r_1, r_2) = \begin{cases} 1 & \text{if } \Delta(r_1, r_2) < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

The constraint violations can be expressed in terms of $w(r_1, r_2)$, $\delta^{\text{S}}(r_1, d)$ and $\delta^{\text{S}}(r_2, d)$ by introducing a binary variable Y_{r_1,r_2} and the following equation

$$w(r_1, r_2) + \delta^{\text{S}}(r_1, d) + \delta^{\text{S}}(r_2, d) + 2Y_{r_1,r_2} = 0 \tag{1.10}$$

[20] Miller, 1995.

Finally, we introduce an indicator variable $I^{(\mathrm{S})}_{w,r_1,r_2}$ to denote whether the above equation holds or not, and obtain the optimization function

$$\min \sum_{\substack{r_1,r_2 \in R \\ r_1 \neq r_2}} \sum_{\mathrm{S} \in \{\mathrm{B},\mathrm{H}\}} I^{(\mathrm{S})}_{w,r_1,r_2} \Delta(r_1, r_2) \tag{1.11}$$

Note that d is implicit (and fixed) in eq. 1.11, *i.e.* our ILP formulation makes inferences on a *per-disease* basis. Further, the optimization function is subject to the following **hard constraints**:

$$\forall r \in R_{d^+}, \quad \delta^{\mathrm{B}}(r, d) = 1 \tag{1.12}$$

$$\forall r \in R_{d^-}, \quad \delta^{\mathrm{H}}(r, d) = 1 \tag{1.13}$$

$$\forall r \in R, \quad \delta^{\mathrm{B}}(r, d) + \delta^{\mathrm{H}}(r, d) \leq 1 \tag{1.14}$$

$$\forall r_1, r_2 \in R, r_1 \equiv r_2, \quad \delta^{\mathrm{S}}(r_1, d) = \delta^{\mathrm{S}}(r_2, d) \tag{1.15}$$

$$\forall d_1, d_2 \in D, d_1 \equiv d_2, \quad \delta^{\mathrm{S}}(r, d_1) = \delta^{\mathrm{S}}(r, d_2) \tag{1.16}$$

Equations (1.12) and (1.13) impose that the relation label of a drug-disease pair whose super-relation has been obtained as prior knowledge *must* remain unchanged. Equation (1.14) requires each pair to acquire at most one super-relation label. Finally, equations (1.15) and (1.16) enforce that synonymous drugs and diseases *must* have identical relation labels.

Defining ϵ: what is "similar enough"?

Our constraint optimization formulation depends on the value of ϵ to decide whether two drugs are deemed similar or not. The constraints imposed in the ILP framework depends on this value. A proper choice for ϵ is therefore fundamentally important.

Since our ILP formulation handles one disease at a time, the super-relation labels are well-defined, *i.e.* given a disease, the drug-to-relation function is not one-to-many. Thus, from the set of drugs in prior knowledge, we constructed the set of all drug-drug pairs with identical super-relation labels. We then computed the distance $\Delta(r_1, r_2)$ for each pair (r_1, r_2) in this set, and selected the top k closest pairs. ϵ was then defined as the average distance between these pairs. In doing so, we manage to discard outliers and obtain a value that is representative of the "typical" similarity between two drugs that bear the same relation with the given disease.

1.6 Experimental Results

We use the UMLS meta-thesaurus to obtain our seed set of relations (*i.e.* the "prior knowledge"). As described in our methodology (Sec. 1.4.1), we derive *beneficial* and *harmful super*-relations from the UMLS relations. These super-relations and the fine-grained UMLS constituents were presented in Table 1.1. In this section of our proposal, we present the experiments and the results obtained for the beneficial drug-disease relations. The experiments on the discovery of harmful relations, *i.e.* drugs that may exacerbate or cause a disease or symptom, is a work in progress. For our experiments, we chose a random set of 10 diseases such that each disease had at least 10 drugs that were known (*i.e.* already present in the UMLS knowledge base) to treat it. These drug-disease relations served as our prior knowledge.

In the remainder of this section, we strive to determine the efficacy of our relation inference methods in finding *new* drugs for therapeutic purposes. To that end, we first describe a supervised

classifier for sentence-level relation extraction, which serves as the baseline. We then present a comparison of the baseline results against the two global inference formulations based on our *latent pathway* model: (i) the *maximum likelihood inference* method (see Sec. 1.5.1), and (ii) the *integer linear programming* formulation (see Sec. 1.5.2).

1.6.1 Sentence-level supervised relation extraction

As our baseline, we built a (distantly) supervised classifier to perform sentence-level relation extraction. To obtain the training data, a set of sentences from the PMC corpus were annotated via the standard distant-supervision paradigm: for each drug-disease pair (r, d) marked as "beneficial" in the prior knowledge, sentences that mentioned both r and d were labeled as positive instances. Likewise, if a sentence contained a drug-disease pair marked "harmful", it was labeled as a negative instance. A little over 57% of the sentences were positive, and the remaining were negative – thus yielding a fairly balanced dataset for the classifier.

This baseline adopts the approach taken by some prior work that has explored supervised relation classification using lexico-syntactic features. These include the methods presented for general domains by Zhang (2004) and GuoDong et al. (2005), and for the biomedical domain by Liu, Shi, and Sarkar (2007), among others. We trained a linear-kernal support vector machine classifier using LIBSVM[21]. The feature space for this classifier comprised of bag of words and fragments from dependency parse trees representing the predicate-theme constructs.

Further, semantic category information of medical entities was also included. This was done in order to ensure that the classifier had access to all the information used to build our latent pathway models. It was trained using 5-fold cross-validation on the training data. Finally, after using the trained classifier to predict the sentences in test data, the output relations were aggregated and sorted according to their classification probability scores.

1.6.2 Ranking

We evaluate the results manually, having human judges who study each positive labeled pair and annotate the relation as true or false. Unlike some prior work in relation extraction using distant supervision who also perform an automatic evaluation by holding out part of their gold-standard knowledge base (*e.g.*, Mintz et al. (2009)), our work, by design, produces drugs that have no relational information in UMLS. This renders automatic evaluation impossible. Also, given the sheer volume of drug-disease pairs produced by the output of all the methods, a manual evaluation of the entire output is not feasible. Therefore, we followed the manual evaluation approach taken by others like Mintz et al. (2009) and Hoffmann et al. (2011), and ranked the inferred relations in order to produce precision/recall curves.

For the baseline classifier, the classification probability scores are used to rank the output. Similarly, for the maximum likelihood inference, the likelihood computed by eq. 1.5 is used. The ILP formulation, however, does not provide any readily available method to rank its output. Next, we describe how a ranking mechanism is incorporated with ILP.

Ranking the predictions made by the ILP formulation

For each disease d, the output of the ILP formulation is a set of drugs that it predicts as beneficial for d. Recall that we measure (dis)-similarity of the pharmacologic effects of a drug by measuring

[21] Chang and Lin, 2011.

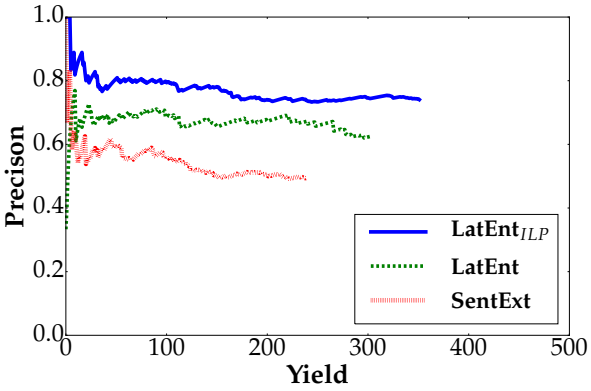

Figure 1.4: Precision *vs.*Yield plot for the three methods of relation extraction/inference: the sentence-level supervised classifier (SentExt), which serves as our baseline, and the two latent pathway models based on maximum likelihood inference (LatEnt) and integer linear programming (Latent$_{ILP}$).

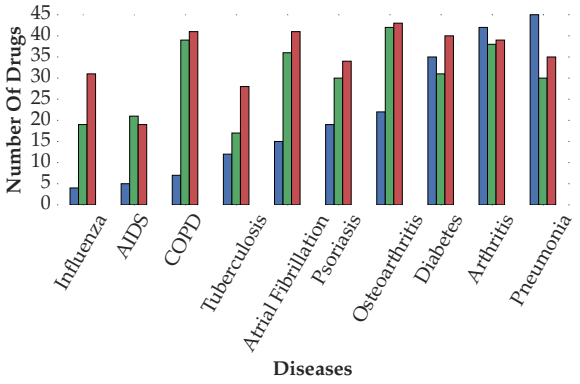

Figure 1.5: Histograms showing the number of new drugs returned by LatEnt formulations against the number of available entires in UMLS. Blue bars show number of drugs in UMLS, green bars show new drugs from the basic LatEnt formulation and red bars show new drugs from LatEnt$_{ILP}$.

the (dis-)similarity of their probability distributions using KL-divergence. Then, ILP uses a cost function based on the divergence measure, as described in Equation 1.7, to arrive at its decisions. Thus, to rank its predictions, we use a scoring function that mimics this cost. Staying true to the spirit of global inference, the score of a new drug r'_j is computed by comparing how similar its effects are to the effects of

(i) the drugs r_i in prior knowledge, *and*

(ii) the new drugs predicted as beneficial by the ILP formulation.

As before, the set of known beneficial drugs is denoted by R_{d^+}. The set of new drugs predicted as beneficial is denoted by R'_{d^+}. We compute the score using the average distance of the drug to the prior drugs and to the other new drugs with the same super-relation label:

$$\text{score}(r'_j) = \frac{1}{2}\left(\sum_{r_i \in R_{d^+}} \frac{D_t(r'_j, r_i)}{|R_{d^+}|} + \frac{\sum_{r'_k \in R'_{d^+}} D_t(r'_j, r'_k)}{|R'_{d^+}|}\right) \tag{1.17}$$

1.6.3 Evaluation

For each disease, we manually evaluated the top 50 drugs returned by each method: the baseline classifier, the maximum likelihood inference (MLI) method, and ILP. Figure 1.4 shows the precision/yield plots for all three. The results show that both the formulations using the latent pathway model clearly outperform the sentence-level relation extraction baseline, even though the feature spaces are identical. The simple MLI formulation, denoted LatEnt, increases overall yield by more than 25% and at a higher precision – 0.62 against 0.49 of the baseline. The more sophisticated ILP-based formulation of the laten pathway model (LatEnt$_{ILP}$), provides the best results at a substantially higher overall precision (0.71) and increases yield by another 16%.

Figure 1.5 shows a comparison of the number of **new** drugs, *i.e.* drugs that are not in a beneficial/harmful relation with the disease in the UMLS knowledge base, identified for each of the ten diseases. In many cases even when the seed knowledge is low (e.g, Influenza and AIDS), formulations based on the latent pathway model are able to identify many new drugs. The bars

Relation Type	Percentage	Description
Therapeutic	24%	Direct expression of therapeutic action on the disease. *e.g.* `prevent`(r, d)
Indirect therapeutic	32%	Expression of therapeutic action on a condition related to the disease or a symptom of the disease. *e.g.* `manage`(r, "pain")
Pharmacologic effect	44%	Expression of targeted pharmacologic effect. *e.g.* `impair`(r, "insulin signaling")

Table 1.4: Distribution of the top 50 predicate-theme pairs for COPD, Diabates and Arthritis across three prominent categories.

also show that the overall performance is consistent across diseases. UMLS, however, is well known to be deficient in relational information. We thus chose to evaluate the usefulness of our relation learning method against DrugBank as well. DrugBank is an expert-curated dataset and is one of the most frequently updated repositories in this domain. Moreover, unlike UMLS, DrugBank has an abundance of relational data in structured form. This result is provided in Table 1.3.

1.7 Analysis

Our data for the results described in the previous section comprised of 100,000 PMC abstracts. Of these, 12,138 abstracts mention at least one drug-disease pair from a UMLS relation. In total, these abstracts yield 2,604 unique UMLS relations, and 1,459 of these relations are expressed within a single sentence at least once. The entities in the remaining 1,145 pairs, however, never co-occur in any sentence, underscoring the need for relation *inference* instead of extraction solely based on lexico-syntactic features.

Disease	Number of *beneficial* drugs not in DrugBank
Influenza	14
Atrial Fibrillation	7
Arthritis	9
Pneumonia	16
Oseoarthritis	18
Psoriasis	15
Tuberculosis	11
AIDS	9
Diabetes	15
COPD	17

Table 1.3: For each disease evaluates, this table shows the number of drugs that were not indicated for it in DrugBank, but were deemed beneficial by human judges. For these 10 diseases, a total of 141 drugs were identified by the ILP formulation as beneficial.

The latent pathway model extracted a total of 112,963 unique predicate-theme pairs for the drugs in prior knowledge. While a large percent of these pairs occur very few times, there are 19,830 pairs that occur three or more times in the abstracts. We analyzed the top 50 predicate-theme pairs for three randomly chosen diseases: COPD, Diabetes, and Arthritis. We found that more than 70% of the candidates expressed relations. These can be broadly categorized into (i) a direct expression of therapy, (ii) an indirect expression of therapy where the treatment is of a condition or symptom related to the disease (*e.g.*, treating "insulin resistance" benefits patients suffering from rheumatoid arthritis), or (iii) expression of targeted pharmacological effects (*e.g.*, a drug lowers blood glucose levels).

1.7.1 Error Analysis

While the global inference formulation using ILP was able to extract relations across sentences and documents with reasonably high performance, there were a few distinct types of errors that led it to incorrect conclusions about drug-disease relations. These are due to the use of MetaMap and UMLS as well as the complex use of language in this domain. We list these main error types below:

1. **Entity recognition:** Biomedical literature is a particularly difficult domain for accurate entity recognition. Several issues such as the presence of large compound nouns (*e.g.*"mitogen activated protein kinase pathway"), incorrect POS tagging (*e.g.*"fasting plasma glucose"), etc. are responsible for this. Further, many entities are long compound nouns comprising of shorter compound nouns that are relevant entities in their own right. The use of UMLS semantic types, too, adds to this type or error. A few entities that are too broad to be useful in our relation extraction problem have the same semantic type as very precise objects. For instance, the very phrase "pharmacological substance" is tagged with the semantic type "pharmacologic substance".

2. **Decoupled compound nouns:** Our work uses extremely simple features, *viz.* predicates and themes in dependency paths, to establish relational connections between entities. This leads to incorrect predicate extraction in case a compound noun is decoupled (*e.g.*"resistance to insulin" instead of "insulin resistance").

3. **Negation:** A particularly difficult aspect of capturing the effect of one entity on another is the complex use of hedging and negation. Instead of direct statements involving 'no', 'not', etc., we often encountered negation expressed by predicates that had a negative connotation in the given context. Phrases such as "slows down the progress of", or "did not contribute to the suppression of" are prime examples of this phenomenon.

4. **Hypothetical Statements:** Many research articles start by saying that their work investigates *whether or not* some medical phenomenon is true. Treatment effects extracted from such sentences also contribute to the noise.

5. **Pathogen Themes:** In some diseases, particularly those that are caused by microorganisms, many documents talk in great detail about the effect of a drug on that pathogen. On these diseases, the themes being affected are components of the pathogen. This leads to a confusing scenario where a drug may be beneficial because it destroys a protein in the pathogen, and our inference procedure mistakenly identifies it with protein destruction in the human host.

1.8 Summary

A visible advantage of the latent pathway model is that it is able to discover not just new treatments that are similar to prior knowledge, but also treatments that have, compared to prior knowledge, a significantly different way of treating a disease. The global inference based on themes enables the system to identify drugs that have *some* pharmacologic effects in common with the drugs in prior knowledge. Often, these new drugs also have other effects not associated with any drug in prior knowledge. The labels on these new drugs further propagate to other drugs that also share these effects. In particular, we observed this propagation in diabetes, where an where an entirely new class of drugs called "sodium glucose co-transporter 2", was discovered from biomedical literature

in spite of no drug from that category being present in the UMLS prior knowledge. It is worth noting that neither the supervised classifier nor the one-step maximum likelihood inference was able to discover these.

In this work, we studied the problem of identifying beneficial drugs for diseases and proposed a method that can extract them by analyzing biomedical abstracts. We proposed a novel latent pathway model that is able to discover relations by extracting information about the pharmacodynamics of a drug. Further, we formulated an ILP model that leverages consistency constraints to perform global inference. Our evaluation showed that this approach achieves substantial improvements over a supervised sentence-level classifier baseline. Our analysis showed that the simple latent pathway model captures useful knowledge, which can be further improved via human curation or through improved modeling.

Chapter 2

Recommending Diagnostic Tests for Identification of Adverse Drug Events

---

An *adverse drug event* (ADE) is an undesired reaction experienced due to use, misuse or discontinuation of medications. Several studies have reported that among the adult population, over 12% of emergency room (ER) visits are caused by ADEs (Capuano et al., 2004; Trifiro et al., 2005; Zed et al., 2008). Safety and quality of patient healthcare are strengthened when a medical problem caused by a drug is promptly and correctly identified. Evidence of an ADE based on a patient's clinical symptoms thus provides an important data point for clinical decision making. However, as there exist way too many drugs, physicians cannot be expected to have memorized all possible ADEs associated with them. Often this is not a problem due to the availability of electronic pharmaceutical databases like Lexicomp[1] or Micromedex[2] that provide extensive information about a wide range of drugs, including their adverse effects. Since this information is provided in the form of narrative texts, to assess the likelihood of ADEs, physicians manually look up these databases. This manual lookup and review is a time-consuming process, prone to lapsed vigilance, and often brings about a failure to order appropriate diagnostic tests (Gandhi et al., 2006).

Diagnostic tests are a critical component of diagnostics because while some symptoms may be observable, many others can only be confirmed by laboratory tests. And even though laboratory testing is the single highest-volume medical activity driving clinical decision making, the process of ordering diagnostic tests and acting upon them remains vulnerable to errors (Singh, 2013; Zhi et al., 2013).

In this chapter, we propose a clinical decision support (CDS) system which automatically *pushes* laboratory test suggestions to confirm (or invalidate) potential adverse effects of a patient's drug regimen. Our application exploits natural language information provided in online pharmaceutical databases. To this end, we use natural language processing (NLP) and template-based techniques to extract three types of information:

$\mathbf{I}_1$. potential adverse effects of a drug.
$\mathbf{I}_2$. observable symptoms associates with medical problems.
$\mathbf{I}_3$. medical problems identified by abnormal values in laboratory test results.

Based on this extracted knowledge, on one hand we map symptoms to medical problems, and the

[1] https://online.lexi.com/
[2] http://micromedex.com/

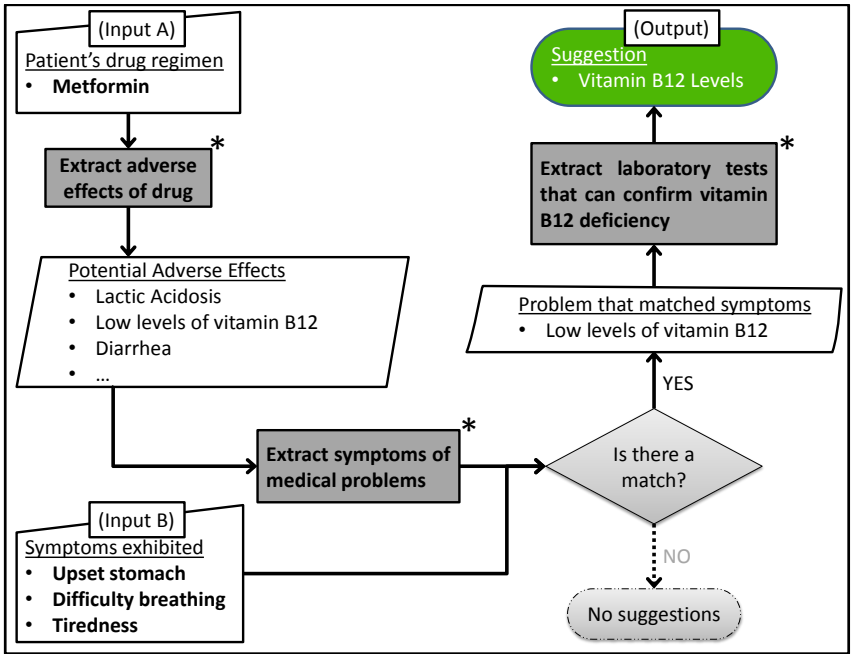

Figure 2.1: A use-case scenario: patient taking metformin (Input A) and exhibiting certain symptoms (Input B). Symptoms are matched to likely adverse effects, and a diagnostic test to confirm this possiblity is suggested (Output). The information extraction steps corresponding to $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ are marked *.

problems to laboratory test results, while on the other hand we map drugs to their adverse effects. Fig. 2.1 shows a use-case scenario where a patient has been prescribed the drug *metformin*, and is experiencing the following symptoms: (a) an upset stomach, (b) difficulty in breathing and (c) overall tiredness.

For information extraction, we use shallow parsing and pattern matching techniques to extract relevant text from available natural language databases. Further, we adapt MetaMap (Aronson, 2001), a biomedical named entity recognizer, to extract medical terms. The three types of information $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ are obtained by employing these steps on three separate *semi-structured* databases, *viz.* Drugs.com[3], the MedlinePlus encyclopedia[4] and Rush University Medical Center's health encyclopedia[5], respectively. In many cases, the output consists of more than one suggestion. This is due to multiple disorders partially or wholly matching the patient's symptoms, as well as multiple diagnostic tests having the ability to identify a medical problem. Providing a clinician with all such suggestions can lead to what is known as *alert fatigue* – the desensitization towards alerts as a consequence of an excessive number of alerts being raised (Koppel et al., 2008). We thus employ similarity resolution to filter out spurious suggestions, and then compute the relevance of each suggestion to finally provide the clinician with a single, short, ranked list.

The remainder of this chapter consists of an overview of our method in 2.1, followed by a description of the datasets and the information extraction steps in 2.2. The similarity resolution and ranking processes are explained in Sec. 2.4 before presenting our experimental results in Sec. 2.5. Finally, we discuss the related work in this area before concluding the chapter.

[3] http://www.drugs.com/sfx/
[4] https://www.nlm.nih.gov/medlineplus/
[5] http://health.rush.edu/HealthInformation. Accessed: Nov 14, 2013

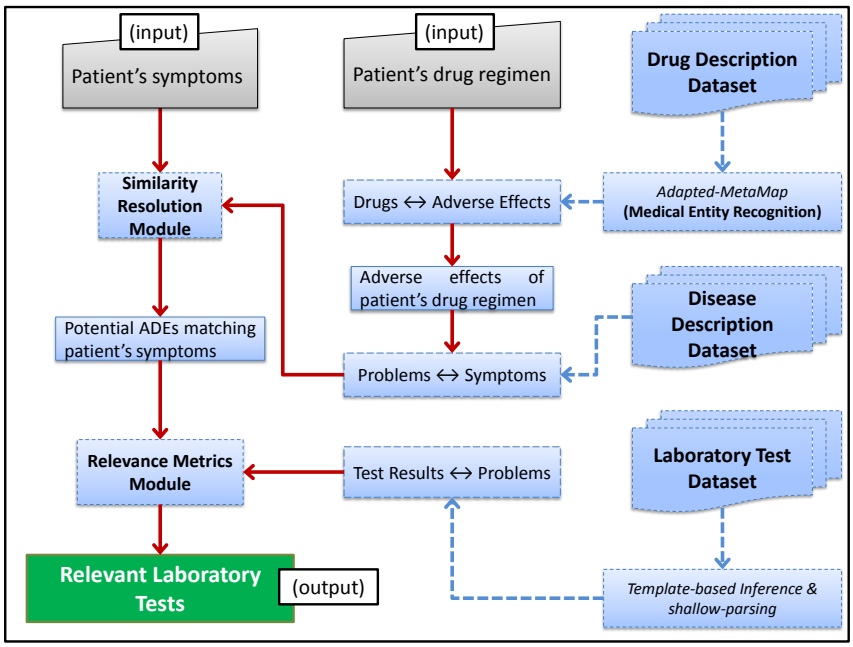

Figure 2.2: Overview of the automated diagnostic test recommendation process. Offline datasets, modules and processes are shown in blue dashes. Information corresponding to $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ are represented as bidirectional maps ($\leftrightarrow$). Segments of the process flow that are input-dependent are shown in solid red arrows.

2.1 Method Overview

We use three separate medical knowledge repositories to automatically suggest diagnostic tests that identify potential ADEs. The first, a drug description dataset, is used to extract the adverse effects of drugs. The second, a laboratory test dataset, is used to ascertain the problems indicated by abnormal test results. Finally, a third dataset contains descriptions of various medical conditions. We use it to map diseases and disorders to their clinical symptoms. The overview of our approach is presented in Fig. 2.2.

Information extraction from these datasets involves the use of MetaMap. It extracts medical terms from natural language data and assigns them semantic types defined by the Unified Medical Language System (UMLS) (Lindberg, Humphreys, and McCray, 1993). For the purposes of this work, we combine several semantic types into two categories, *medical conditions* and *drugs*, as shown in Table 2.1. The distinction between our semantic categorization and that followed by MetaMap is further explained in Sec. 2.2.1.

Given a patient's list of medications and an optional list of exhibited symptoms, our application identifies a list of medical conditions $\{c_1, \ldots, c_k\}$ fitting two criteria:

(i) Each c_i is a potential adverse effect of at least one of the drugs being taken by the patient.

(ii) The symptoms of each p_i match at least some of these exhibited symptoms.

For the second criterion, i.e. matching symptoms, synonyms are resolved so that equivalent symptoms are identified. For example, if the disease description dataset provides "fatigue" as a symptom of an ADE while the input symptoms contain "tiredness", the *synonym resolution module* desribed in Sec. 2.3 identifies them as equivalent.

Finally, we perform a relevance ranking of laboratory tests based on the similarity between (a) medical problems identified by test result, and (b) the list of problems $\{c_1, \ldots, c_k\}$ obtained above.

Category	UMLS Semantic Types
Medical Condition	Disease or Syndrome; Sign or Symptom; Body System; Laboratory or Test Result; Mental or Behavioral Dysfunction; Cellular or Molecular Dysfunction; Mental Process; Individual Behavior; Neoplastic Process; Acquired Abnormality; Anatomical Abnormality; Congenital Abnormality
Drug	Organic Chemical; Biologically Active Substance; Pharmacologic Substance; Amino Acid, Peptide or Protein; Steroid; Clinical Drug

Table 2.1: Categories and their constituent UMLS semantic types.

2.2 Information Extraction

This application required the extraction of three types of information in order to map (a) drugs to their adverse effects, (b)medical problems to their clinical symptoms, and (c) laboratory tests to the medical problems identified by them. These mappings are obtained by extracting information from the drug description dataset, the disease description dataset, and the laboratory test dataset, respectively. In this section, we present the details of these repositories along with the extraction processes employed to distill structured information from them.

2.2.1 The drug description knowledge base

For this work, we used a publicly available web-repository, Drugs.com, to obtain relevant information about drugs typically used in patient care settings like the ER. This repository consists of $5,856$ unique drugs. It is a *semi-structured* dataset, *i.e.* the different types of information are pre-labeled. All the information about a drug is presented under sections with labels such as "side effects", "dosage", "warnings", etc. Within each section, however, the information is presented in descriptive, and often complex, natural language text. To extract the relevant information from these texts, it is critical to correctly identify the relevant medical entity mentions.

Medical Entity Recognition (MER) is the first step of this process. A medical entity is a particular instance of a medical concept or category. For example, the drug "metformin" is an instance of a pharmacologic substance. Recognizing such entities requires first, detection of their mentions in the text, and second, identifying their semantic category. Next, we describe how we tailored MetaMap for successful MER in this application.

Tailoring MetaMap

In the scope of this work, we used the UMLS semantic types[6] to broadly categorize entities into the types that are of interest in our healthcare application: *drugs* and *medical conditions*. Table 2.1 shows the complete list of the constituent UMLS semantic types that make up these two broad categories. This coarse categorization is needed because MetaMap simply labels phrases with a list of UMLS semantic types. These labels, however, are too fine-grained for our purpose. Moreover, the labeling is not precise enough to be readily used in a system built for real patient care settings.

MetaMap operates by first splitting a text into its constituent sentences. It then runs the MedPost/SKR part-of-speech tagger (Smith, Rindflesch, and Wilbur, 2004), a tagger built for biomedical

[6] The complete list of 135 semantic types is available at https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

Phrase	UMLS Semantic Type Label	MetaMap Score
"Subnormal vitamin B12 levels have been reported, and may result in anemia or neuropathy."		
Subnormal vitamin B12 levels	Laboratory or Test Result	906
Anemia	Disease or Syndrome	746
Neuropathy	Disease or Syndrome	1000
"In vivo evaluation of Soman and Tabun Inhibited Acetylcholinesterase."		
In vivo evaluation of Soman	Health Care Activity	760
	Spatial Concept	640
	Pharmacologic Substance	593
Tabun	Organophosphorous Compound	1000
AcetylCholinesterase	Amino Acid, Peptide, or Protein, Enzyme	1000

Table 2.2: Entity extraction using MetaMap (without any tailoring), illustrating two kinds of errors: (i) perfect matches like "anemia" often have low scores, and (ii) relevant entities like "Soman" may get incorporated into larger phrases and get low scores. Further, phrases like "subnormal vitamin B12 levels" are, for the purposes of our application, adverse effects, beyond being test results. Such application-specific labeling is, of course, not readily available.

language, and a shallow parser that further chunks each sentence into phrases. Following this, it generates variants of the phrase, which includes not just linguistic variations, but also any subsequence of words in the phrase that appears in the UMLS lexicon[7]. Finally, MetaMap ranks all the mappings and returns a scored list of medical entities. The ranking function is based on four simple metrics that measure the extent of linguistic variation and overlap between the actual phrase in the text and the mapped entity in the lexicon. Even though in his detailed report, Aronson (2006) describes some heuristics to focus on correctness rather than breadth, this process generates too many mappings for each phrase – including a high number of incorrect entities. Further, it is often the case that the top ranked entity in the returned list is not suitable for our application.

The labels readily offered by MetaMap are not conducive for our application for three primary reasons. First, labeling errors often arise from incorrect phrase chunking. Second, we observed that in many cases, even a perfect lexical match did not yield a high score in MetaMap, and third, for the purposes of our application to work as a CDS system, we need to incorporate additional semantic types into our coarse-grained "super"-categories, even though ordinarily, the labels assigned to them by MetaMap would suffice. The examples shown in Table 2.2 serve as an illustration of these errors. Thus, to improve the precision of MER with CDS as the goal, we made the following revisions to MetaMap's labeling process:

(i) Perform shallow parsing with Genia (Tsuruoka et al., 2005). Even though both the MedPost/SKR and Genia taggers are tailored to the biomedical domain, we found that using the latter lead to phrase generation at a finer granularity, and as a result, considerably reduced errors of the first type.

(ii) Filter mappings based on an empirically determined score threshold.

(iii) Filter mappings by imposing additional restrictions on semantic types. For example, the semantic type "Element, Ion or Isotope" is a child of the UMLS category "Chemicals and Drugs". This labeling, while relevant for information extraction from research literature (as discussed in Chapter 1), hurts the precision of a typical CDS while offering only marginal improvements in recall.

(iv) Include some additional semantic types in the super-categories (see Table 2.1).

[7] *UMLS® Reference Manual [Internet]* (2009, Ch. 6). Available: http://www.ncbi.nlm.nih.gov/books/NBK9680/

Laboratory Test Knowledge Base Vitamin B12 Level Test	**Disease Description Knowledge Base** Vitamin B12 deficiency
Vitamin B12 Level Definition ... Normal Values ... Values of less than 200 pg/mL are a sign of a vitamin B12 deficiency. Causes of vitamin B12 deficiency include diseases that cause malabsorption (*e.g.* Celiac disease and Crohn's disease), not enough vitamin B12 in diet, ... Conditions that can increase vitamin B12 levels include liver disease, ...	Symptoms can include: • Diarrhea or constipation • Fatigue, lack of energy • Light-headedness when standing up • Loss of appetite • Pale skin • Problems concentrating • Shortness of breath, mostly during exercise • Swollen, red tongue or bleeding gums

Table 2.3: Typical examples of semi-structured information as available in the laboratory test and disease description knowledge bases. For information extraction, this is simpler than unstructured knowledge bases, and template-based inference methods suffice. Linguistic cues of such templates (*e.g.* "causes of") are underlined.

In Sec. 2.5, where we present our experimental results, we include an extrinsic evaluation of this application-specific adaptation of MetaMap by juxtaposing the performance of our CDS system with the original MetaMap labels against the same system using the labels obtained with the above modifications. Next, we describe information extraction from the laboratory test and disease description datasets.

2.2.2 Shallow Parsing and Template-based Inference

The two remaining knowledge bases (KBs) we use are also semi-structured, but linguistically simpler. As a result, we were able to extract the relevant pieces of information from them using a combination of shallow parsing (to identify phrasal cues) and template-based rules. As our KB for laboratory tests and procedures, we used the Rush University Medical Center's health encyclopedia. It comprises of 603 laboratory tests, where information about each test is provided in semi-structured narratives. In order to determine the observable symptoms associated with medical problems, we use the National Library of Medicine's MedlinePlus encyclopedia as the disease description KB. From this source, we extracted information about medical problems, including the clinical symptoms they manifest. Table. 2.3 shows typical semi-structured templates from these two sources.

Even when the information is unstructured, it is in the relatively simple form of short sentences that serve as data labels, *i.e.* sentences from which the *type* of information can be inferred. To identify them, we chunk sentences into phrases, and perform a simple frequency-analysis. For this analysis, we extract verbal phrases and prepositional phrases that follow a nominalized predicate. Further, we only consider phrases occurring in a left- or right-context window of 4 words around the disease mention in a sentence. They are then lemmatized word-by-word. This step normalizes expressions such as "cause of" and "causes of". Sorting such phrases by their frequency showed that there is little variation in the way symptoms and causes are described in these knowledge bases, thus eliminating the need for

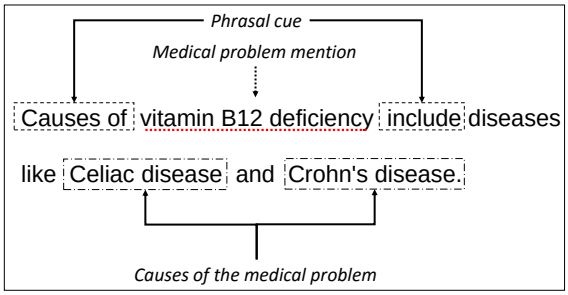

Figure 2.3: Causes of vitamin B_{12} deficiency identified using the template $T = \langle$"`cause of`" CONDITION "`include`" $c_1, c_2, \ldots$ *and* $c_n\rangle$.

anything more complex than designing template-based rules. Figure 2.3 illustrates the use of a simple template for extraction of the potential causes.

We also designed templates that decouple compound nouns to obtain the details of a medical condition. Performing a frequency analysis similar to the one described above, we observed that almost all compound nouns that were mapped by MetaMap to a UMLS semantic type in the "medical condition" category, but also got a (lower) score for the "drug" category, were conditions expressed in terms of a polar qualifier attached to a pharmacological or physiological entity. Common examples included phrases like "*high* blood pressure", "iron *deficiency*", "*excess* vitamin D", etc. These qualifiers, too, varied little across the disease description and laboratory test KBs.

We were thus able to design templates that mapped laboratory tests to symptom descriptions. For example, using a template to identify the qualifier "deficiency" in Fig. 2.3, we were able to (a) extract the core entity vitamin B_{12}, and (b) relate the extracted information, *i.e.* "Celiac disease" and "Crohn's disease", to *low* values of the corresponding laboratory test result. This simple method suffices because laboratory tests typically have two types of abnormal results: (a) results *above* the reference range, and (b) results *below* it. Laboratory tests that return binary results, *i.e.* positive or negative, were also matched using similar templates. Using the shallow-parsing and template-based inference rules, we identified the types of abnormal test results (*e.g.* low/high, positive/negative). Further, we used MetaMap to extract the medical problems associated with such abnormal results. This way, we obtained a directed tripartite graph that maps each abnormal test result to a medical problem, as shown in Fig. 2.4). Some test results, of course, may not indicate any abnormality.

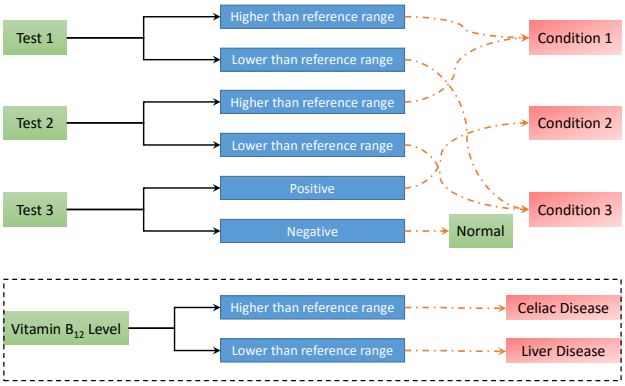

Figure 2.4: Mapping laboratory test results to conditions.

2.3 Similarity Resolution

Even though for this application, we were able to choose some semi-structured knowledge bases where template-based rules were sufficient for accurate extraction of symptoms and laboratory test results, polysemous expressions of symptoms were commonplace. Our initial experiments showed that ignoring such variations had a negative effect on identification of relevant ADEs. However, these KBs mostly only employed word-level polysemy. Thus, instead of attempting to identify all possible paraphrases – a difficult problem in NLP, especially in this domain where semantics are often highly contextual – we tackle this issue at a simpler level by resorting to ontology-based semantic similarity measures.

To this end, we employ one general-purpose ontology, WordNet (Miller, 1995), and a medical vocabulary, MeSH (Medical Subject Headings)[8]. The latter is the National Library of Medicine's comprehensive controlled vocabulary for the biomedical domain wherein detailed category information is available for each term. In both WordNet and MeSH, synonymous terms are identified under a concept-set. Staying in line with the terminology used by WordNet, the remainder of this section will refer to such a set of semantically equivalent terms as a *synset*.

[8] (Rogers, 1963) Available: https://www.nlm.nih.gov/mesh/

Synsets available from WordNet and MeSH are able to directly resolve several equivalent terms, *e.g.* "fatigue" and "tiredness". . Identifying the semantic similarity in a more general sense (*i.e.* going beyond synonyms), however, requires a little more work. Many symptoms are expressed in words or phrases that are used interchangeably. Additionally, many clinical expressions offer a narrower or broader semantic scope than the symptom itself. For example, a patient suffering from a "histamine headache", a MeSH synset, may have her symptom listed only as "headache". Such closely related expressions are often treated by clinicians as equivalent, when actually they are closely related, but distinct entities. These terms often share a *type-of* relation. If e_1 is a type of e_2, it is said that e_1 is a *hyponym* of e_2, and e_2 is a *hypernym* of e_1. Both WordNet and MeSH contain hyponym and hypernym relations. Globally, these relations form a directed acyclic graph. The further away two words are from each other in this graph, the less similar they are. In other words, the shortest path length $l(t_1, t_2)$ between two terms t_1 and t_2 is inversely related to their semantic similarity $\delta(t_1, t_2)$, *i.e.* $\delta(t_1, t_2) \propto 1/l(t_1, t_2)$.

There is a formidable body of prior research delving into similarity metrics based on the path between two concepts in an ontology. With WordNet, in particular, a formulation based directly on the path length was demonstrated by Wu and Palmer (1994). For the purposes of identifying semantically similar terms in our KBs, we follow their approach. Note that unlike their work, our situation dictates that terms be identified as equivalent if they are similar *enough*. Therefore, instead of using the metric itself, we simply experiment with varying path lengths to seek a desirable threshold for such a definition of semantic equivalence. Later in section 2.5, we present the performance of our application with different path length thresholds.

2.4 Avoiding Alert Fatigue

Clinical decision support (CDS) systems aimed toward mitigating the adverse effects of medications have often caused *alert fatigue* among healthcare practitioners (Koppel et al., 2008). This is due mainly to the fact that almost every drug has *some* side effect, no matter how mild. Some of these undesirable effects are caused by a vast majority of drugs. Nausea and dizziness, for example, can be caused by almost any medicine (Scorza et al., 2007). Similarly, there are symptoms that can be manifested by many medical conditions, and are too common to warrant a diagnosis involving laboratory tests. If all such adverse effects are considered, and an application starts suggesting laboratory tests to confirm or invalidate each one of them, it has been observed that such automated alerts soon start getting ignored. In some cases, clinicians were found to be ignoring up to 96% of the alerts (Sijs et al., 2006).

These studies clearly indicate that the tendency of CDS systems to cause alert fatigue greatly diminishes their potential to improve patient safety. We thus present the final step of this application, wherein several potentially useful laboratory tests are ranked according to their relevance to a given set of medical conditions. Our goal is to ensure that the most relevant recommendations appear at the top of this list.

In order to do this, we adopted a two-pronged approach. On one hand, for each test t in the laboratory test KB, we extracted a set of medical conditions C_t, *viz.* the conditions confirmed by abnormal values of t. On the other hand, we extracted all possible adverse effects of the patient's medications. This yielded another set of conditions C_m. These extraction processes are the ones described previously in section 2.2.2. The tests are then ranked in descending order of their Jaccard similarity coefficient (Jaccard, 1901) between the two sets of conditions C_t and C_m, defined as the

size of the intersection divided by the size of the union of the sample sets

$$J(C_t, C_m) = \frac{|C_t \cap C_m|}{|C_t \cup C_m|} \tag{2.1}$$

Thus, a laboratory test that can confirm or discard a medical condition that is indicative of a higher number of a patient's symptoms is given a better rank.

2.5 Experimental Results

Our application is intended to serve as a CDS system in direct patient care settings like an ER – the last stage of the translational pipeline. Based on a patient's drug regimen and symptom manifestations, it automatically suggests relevant laboratory tests that can confirm or invalidate an ADE. There is, however, no gold standard data available for this. Moreover, since the current practice in healthcare facilities of ordering laboratory tests is error-prone (Singh, 2013; Zhi et al., 2013), a comparison against real diagnostic processes will also not be truly indicative of the performance of our application.

In this section, we thus present an evaluation based on a small dataset comprising of the first 40 drugs from the list of top 100 drugs (by sales) available on Drugs.com. For each drug in this set, we manually annotated all possible adverse effects and the laboratory tests (if necessary) that are capable of confirming them. Our evaluation attempts to answer the following questions:

($\mathcal{Q}_1$) Are the diagnostic test recommendations useful if no observable symptoms are present?

($\mathcal{Q}_2$) Are the diagnostic test recommendations capable of confirming/discarding an ADE if observable symptoms are provided?

In the process, we also evaluate our information extraction methodology. In particular, we provide insights on whether or not

($\mathcal{Q}_3$) our tailored approach toward medical entity recognition improves the quality of recommendations over the standard use of MetaMap, and

($\mathcal{Q}_4$) our similarity resolution module improves the quality of recommendations over naïve matching based on keywords and synonyms.

The remainder of this section is devoted to addressing these queries one by one.

($\mathcal{Q}_1$) Suggestions based solely on patient's drug regimen

When suggesting diagnostic tests without being provided any of the patient's symptoms, our application is unaware of the prior likelihood of any ADE. All potential adverse effects are thus treated as equally probable. Therefore, instead of ranking by relevance, we check whether a test suggested by our application has the ability to identify *any* potential adverse effect associated with a given drug. For evaluation, two human judges independently assigned a *relevant* or *irrelevant* label to each laboratory test recommended by the system. For the 40 drugs, a total of 226 tests were suggested, of which 186 were deemed relevant by the judges. The remaining tests were labeled as irrelevant by at least one judge. Additionally, 28 laboratory tests that were found to be relevant by the judges were not suggested by our system.

In our evaluation data, 10 out of the 40 drugs are psychotropic medications. Upon analyzing the results, we observed that in general, test suggestions to identify potential ADEs associated with psychotropic medications performed poorly. The worst results were obtained for the drug Adderall, where only 2 of the 11 suggested tests were judged as relevant. For the other 30 drugs, it was

Evaluation Set		*standard* MetaMap		*tailored* MetaMap	
		Recall	**Precision**	**Recall**	**Precision**
All Drugs	All Symptoms	0.84	0.80	0.87	0.82
Nonpsychotropic Drugs	All Symptoms	0.84	0.81	0.86	0.85
Nonpsychotropic Drugs	*minus symptoms associated with the nervous system*	0.85	0.82	**0.89**	**0.86**

Table 2.4: Laboratory test suggestions based solely on patient's drug regimen, without any symptoms being provided.

noted that suggestions for ADEs involving the nervous system were comparatively less accurate. The performance of our system based only on the patient's drug regimen is shown in Table 2.4, which also attends to $\mathcal{Q}_3$ by demonstrating the improvement achieved by tailoring the medical entity recognition process (see Section 2.2.1).

($\mathcal{Q}_2$) Suggestions based on patient's drug regimen and symptoms

Our second experiment takes into account the scenario where, in addition to the list of medications, we also know a few symptoms exhibited by the patient. We test with each drug twice by providing two sets of symptoms s_1 and s_2 corresponding to different medical problems. For instance, the drug 'metformin' was tested with the two symptom-sets {nausea, vomiting, low blood pressure} and {fatigue, dizziness, dyspnea, diarrhea}. These two sets of symptoms correspond to two potential adverse effects of metformin, *viz.*, 'lactic acidosis' and 'vitamin B_{12} deficiency', respectively. We thus obtain 80 test data points for the 40 drugs. For each input, the output is a list of laboratory test suggestions ranked by the Jaccard similarity coefficient described in Eq. 2.1.

As part of these experiments, we test the similarity resolution step as well, thereby addressing point $\mathcal{Q}_4$. Our results show that resolving similarity leads to a significant improvement in performance when direct hyponymy/hypernymy are incorporated. Longer paths, however, simply result in more tests being suggested, even if they are for ADEs with quite different symptoms. Our application provides a list of laboratory tests that can confirm (or invalidate) a medical condition associated with the patient's symptoms provided that the condition is also a known adverse effect associated with the patient's drug regimen. The output list is ranked by the relevance of the test in identifying a potential ADE associated with the given set of drugs and symptoms. To evaluate such a list, we measure the *mean reciprocal rank* (MRR) (Voorhees et al., 1999), a widely used metric in information retrieval (a detailed survey can be found in Baeza-Yates, Ribeiro-Neto, et al. (1999)) to evaluate systems that provide ranked results instead of a single answer. It is the average of the reciprocal ranks of the output list

$$MRR = \frac{1}{|S_i|}\sum_{k=1}^{|S_i|}\frac{1}{\text{rank}_k} \tag{2.2}$$

where S_i is the input set of drugs and symptoms, and rank_k is the rank of the correct suggestion for the k^{th} test input. The ranking evaluation using this metric is presented in Table 2.5, which includes comparisons between using similarity resolution with different path lengths (see Section 2.3) and not using it at all.

Evaluation Set		Mean Reciprocal Rank				
		¬(SimRes)	SimRes			
			$l = 0$	$l = 1$	$l = 2$	$l = 3$
All Drugs	All Symptoms	0.78	0.83	**0.88**	0.76	0.73
Nonpsychotropic Drugs	All Symptoms	0.80	0.86	**0.89**	0.79	0.74
Nonpsychotropic Drugs	*minus symptoms associated with the nervous system*	0.80	0.85	**0.93**	0.78	0.74

Table 2.5: Laboratory test suggestions based on patient's drug regimen and symptoms, with different similarity resolution thresholds. Results obtained without the use of similarity resolution are indicated by ¬(SimRes). In the subsequent columns l denotes the shortest path length between two terms within which they were considered semantically equivalent. E.g., $l = 0$ indicates that only synonyms were considered equivalent.

Error Analysis

There are two aspects of the experimental results that we probed into. First, the distinctly inferior performance upon encountering psychotropic drugs, and second, a further deterioration when coupled with symptoms pertaining to the central nervous system. Analyzing the errors in such cases, we found that our recommendation system is not fully capable of interpreting terms that tend to be rather subjective. These terms are particularly prevalent in the descriptions of psychological conditions. The symptoms of these maladies are often vaguely defined in terms of mental faculties like "lack of concentration", "abnormal behavior", etc. Symptoms associated with the central nervous system are also often at least partly psychological in nature. Examples include "nightmares", "laziness", "mood changes", etc.

The simple ontology-based similarity resolution that we perform is not adequate for these cases. Many terms that occur in these disease descriptions often have several synsets with non-medical semantics. Using WordNet, therefore, hurts the precision of our system. Another important reason for the relatively poorer performance with psychotropic drugs and psychological symptoms is that a majority of such symptoms are detected by human judges through conversations, not measurements based on physical tests. In many cases, the symptoms are very similar to normal behavior in human beings, but may be present in a far greater degree in patients who actually suffer from a psychological condition. A well-known example is social anxiety disorder, whose symptoms include anxiety and fear of getting judged by those around us. To some extent, these reactions affect a vast majority of people. However, only through detailed psychological evaluations can terms like "anxiety" and "fear" be treated as clinical symptoms, since there are few, if any, laboratory tests to identify such cases.

2.6 Related Work

The topic of pharmacovigilance, defined as the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem[9], has received substantial research attention in the medical literature (*e.g.* Beuscart, Hackl, and Nøhr (2009) and Friedman (2009), among others). Computer-aided approaches in this area can be divided into two orthogonal categories: (a) extraction of new ADEs from population data, and (b) detection or identification of already known ADEs in patients.

[9] World Health Organization, 2006.

Extraction of ADEs based on population studies

This body of work aims to discover previously unknown ADEs from population-based historical clinical data – primarily a pharmacoepidemiological approach. Much of this work is based on mining biomedical research literature (Agbabiaka, Savovic, and Ernst, 2008; Wang et al., 2011; Gurulingappa, Mateen-Rajput, Toldo, et al., 2012; Théophile et al., 2012). The main drawback is that their supervised learning requires large amounts of annotated training data (Gurulingappa, Mateen-Rajput, Toldo, et al., 2012; Segura-Bedmar, Martínez, and Sánchez-Cisneros, 2011; Segura-Bedmar, Martinez, and Pablo-Sanchez, 2011), or else suffers low accuracy (Théophile et al., 2012; Shetty and Dalal, 2011; Percha, Garten, and Altman, 2012). Getting large amounts of annotated text is, of course, inherently expensive. The problem we addressed in this chapter, however, is not the discovery but the identification of known ADEs that are described in narrative form in existing pharmaceutical databases.

Identification of ADEs in patients

Prior research tackling the identification problem has focused on creating ADE alerts in patient data. In one line of work, alerts are formulated as rules that get triggered whenever signs and symptoms in patient data satisfy the conditions attached to these rules. The conditions are built with medical terms in the patient's symptoms and ADE descriptions. These terms are drawn from discharge summaries (Melton and Hripcsak, 2005; Wang et al., 2009), ADE reports (Botsis et al., 2011; Tatonetti, Fernald, and Altman, 2011) and ambulatory care notes (Cantor, Feldman, and Triola, 2007). NLP tools based on MetaMap, like MedLEE (Friedman, 2000), have also been used to extract them.

The rules in these approaches vary in complexity. Some, like Honigman et al. (2001), Cantor, Feldman, and Triola (2007), and Haug et al. (2007), have used simple rules based on keywords or short linguistic patterns to trigger alerts. As noted by Murff et al. (2003), such triggers do not achieve a desirable level of accuracy. Additionally, the rules are learned from a set of documents obtained from a small number of hospitals, often one. In other words the data set from which the rules are derived is far from being comprehensive. Not surprisingly, such systems report low accuracy even when employing more sophisticated NLP tools to extract relevant medical terms (Melton and Hripcsak, 2005; Wang et al., 2009). More complex decision rules have also been used which combine keywords and laboratory findings (Botsis et al., 2011; Seger et al., 2005; Gandhi et al., 2010; Tatonetti, Fernald, and Altman, 2011). However, in this body of work, the rules are manually curated. Notable exceptions include the PSIP project (Beuscart, McNair, Brender, et al., 2009), which performs more complex semantic mining. Their approach does not, however, extend to external natural language databases.

Identification using external knowledge sources

Some recent work has explored the utility of available semi-structured knowledge sources. Notable in this class, for detection and prevention of ADEs, is the use of RxNorm[10] by Izquierdo-Garcia and Escobar-Rodriguez (2012), Smithburger, Kane-Gill, and Seybert (2012), and Tsai et al. (2013), among others. These approaches continue to offer shallow coverage, though, due to their focus on very specific drugs (*e.g.* Izquierdo-Garcia and Escobar-Rodriguez (2012) and Tsai et al. (2013)) or

[10] http://www.nlm.nih.gov/research/umls/rxnorm/

because they continue to build rules manually (*e.g.* Smithburger, Kane-Gill, and Seybert (2012)). More importantly, this body of work does not exploit laboratory test data.

In summary, despite substantial research in ADE detection methods, gaps remain. Our approach attempts to fill this gap with three salient characteristics:

(a) extensive utilization of available pharmaceutical databases,
(b) use of automated measures that provide coverage over a large set of drugs, and
(c) combining diagnostic test information with ADE information to aid clinical decision support.

2.7 Summary

In this chapter, we presented an approach for automatically suggesting diagnostic tests as a contribution to the last phase of the translational research pipeline. We presented the details of this clinical decision support system, and in particular, demonstrated its practical use in the identification of potential ADEs. As a part of this process, we perform medical entity recognition and template-based information extraction from available datasets about adverse drug effects, laboratory tests and symptoms of various medical conditions.

A key observation that we would like to make at this point is that in cognizance of the translational research pipeline, once new knowledge from biomedical research literature is incorporated into semi-structured knowledge bases where relations between medical entities are explicitly mentioned in relatively simpler language, building healthcare applications with tangible benefits becomes a much simpler process. Thus, the work presented in chapter 1 can be seen as having an impact through the incorporation of the extracted relations into existing KBs, and then building applications such as the one described in this chapter. In the next chapter of this proposal, we present another application that builds on relational information distilled from biomedical literature, and *translates* it to the proverbial "bedside", in an emergency room setting.

Chapter 3

Identification, Attribution and Ranking of Adverse Drug Events

---

This third and final chapter of the proposal presents our second contribution towards the second half of the translational research pipeline: bringing the outcome of research investigations into clinical practice. We present a clinical decision support (CDS) system that makes extensive use of the kind of *relational knowledge* we learned in Chapter 1, and automates the identification and attribution of adverse drug events (ADEs) in an emergency room (ER) setting. In Chapter 2, we focused on one error-prone area of healthcare, *viz.*, incorporating laboratory tests in the diagnostic process. In particular, we noted that a recommendation system such as the one we presented, can provide evidence in support of ADE diagnoses. In this chapter, our focus is entirely on ADEs. As the title suggests, the work we present here can be viewed as a three-pronged approach comprising of

(a) **identification** of potential adverse effects based on the patient's drug regimen,
(b) **attribution** of the patient's symptoms and complaints to such ADEs, and
(c) **ranking** of such attributions based on the strength of their association with the drugs and symptoms.

3.1 Motivation

An ADE is a undesirable reaction experienced due to the use, misuse or discontinuation of medications. It is an alarming truth that a high number of hospital ER visits are due to them. A widely cited study conducted by Null et al. (2005) reported $783,936$ iatrogenic[1] fatalities at an estimated financial cost of $\$282$ billion, and a total of 2.2 million in-hospital adverse drug events. A sizeable body of more recent work has also repeatedly pointed out this phenomenon, with notable examples including Trifiro et al. (2005), Zed et al. (2008), and Jayarama, Shiju, and Prabahakar (2012).

Information about ADEs is often available in semi-structured drug databases. For attribution of the patient's symptoms and complaints to ADEs, it is necessary for physicians to manually review these narratives in light of the patient's medications. However, over 40% of such cases are overlooked by emergency physicians (Hohl et al., 2010; Roulet et al., 2014). Given that most

[1] *Iatrogenesis* is defined as "inadvertent and preventable induction of disease or complications by the medical treatment or procedures of a physician or surgeon". From: http://www.merriam-webster.com/medical/iatrogenesis

ERs are overcrowded, such lapses in patient care are potentially an outcome of increased workload (Trzeciak and Rivers, 2003; Olshaker and Rathlev, 2006; Weissman et al., 2007; Collis, 2010). Overcrowding also means that spending more time on a single patient may not be feasible. Indeed, it has been observed that under such circumstances, physicians increasingly restrict themselves to only those questions that can be instantly answered (Ramos, Linscheid, and Schafer, 2003). Unfortunately, the current practice of ADE attribution requires a physician to manually read through narrative texts in online pharmaceutical databases such as Lexicomp[2] or Micromedex[3]. As a result, even while missing nearly half of the ADEs, they already suffer from what has been called the "$4,000$ click syndrome" – spending much of their time with electronic records rather than in direct patient care (Hill, Sears, and Melanson, 2013).

The ER setting thus warrants an evidence-based CDS system that automatically detects possible ADEs and instantly *pushes* such diagnostic suggestions to the physician, making it an instantaneous process not requiring any clicks. Such a system will not only improve patient safety by identifying ADEs, but also save a significant amount of time in crowded ERs by allowing physicians to quickly determine an important step in their medical diagnosis. This second aspect is particularly important due to the exponentially high number of potential adverse effects that can arise from drug interactions. In research as well as in clinical practice, drug interactions have largely been considered as pairwise events involving two drugs (Horn, Hansten, and Chan, 2007), and higher-order interactions have rarely been studied (Mannheimer, 2009). Due to the lack of available resources on higher-order drug interactions, they are beyond the scope of this proposal. Even with this restriction to pairwise interactions, a patient taking 12 drugs requires the physician to check up to 78 possible cases (12 drugs and 66 potentially interacting pairs). Even if s/he spends only a minute per potential complication to look up the pharmaceutical database and skim through the adverse effect narrative, it would be far beyond the feasible amount of time for this task, especially in ERs, which more often than not, tend to heavy traffic.

3.2 Related Work

Pharmacovigilance, the discipline pertaining to the detection, assessment and prevention of drug-related adverse events, has been a topic of significant interest in the healthcare community. Computer-aided approaches aimed at ADE prevention in patients has focused on creating tools and services that raise alerts by prompting clinicians about potential ADEs. A vast majority of these, however, are prevention systems designed to issue an alert at the moment of prescribing new medication. These systems have been presented by Galanter, Didomenico, and Polikaitis (2005), Schedlbauer et al. (2009), and Jha et al. (2009), among others. Even though such CDS implementations are aimed at preventing adverse events, they have not always improved patient safety (Gurwitz et al., 2008; Strom et al., 2010). The lack of a pervasive and tangible improvement is due to two seemingly conflicting factors: *coverage* and *alert fatigue*.

Coverage versus Alert Fatigue

In one line of work, alerts are rule-based, where the conditions for raising an alert are built from medical terms in a patient's symptoms and ADE descriptions. These terms are drawn from discharge summaries (Melton and Hripcsak, 2005; Botsis et al., 2011), ambulatory notes (Cantor, Feldman,

[2] https://online.lexi.com/
[3] http://micromedex.com/

and Triola, 2007) or ADE reports (Wang et al., 2009; Tatonetti, Fernald, and Altman, 2011). Entity extraction systems based on Metamap (*e.g.* MedLEE (Friedman, 2000)) have also been used. But because the rules are manually curated, such approaches suffer from low coverage. On the other hand, rule-based trigger systems that have attempted to generalize in order to improve coverage tend to raise too many alerts, thereby causing alert fatigue (Rozich, Haraden, and Resar, 2003; Classen et al., 2011; Baseman et al., 2013). In some cases, clinicians were found to be ignoring upto 96% of the raised alerts (Sijs et al., 2006). Keeping these issues under consideration, our work adopts a three-step approach:

(a) To ensure maximum coverage, exploit multiple knowledge bases for each task.
(b) To produce relevant ADE diagnosis suggestions, *normalize* entities and filter out ADE attributions not adequately supported by evidence.
(c) Finally, *rank* the remaining suggestions.

Some prior studies have argued that ranking is a more suitable approach compared to filtering (Sijs et al., 2008; Lee et al., 2010), but more recent reports suggest that expert panels recommend filtering as well (Phansalkar et al., 2013). We thus adopt a combination of filtering and ranking. Further, unlike the ranking algorithms in existing CDS systems, our simultaneous use of several types of information allows for the final suggestions to be ranked based on a combination of multiple factors like the severity of an ADE, the likelihood of its occurrence, relevance to patient's symptoms, etc.

Identification of ADEs using external knowledge sources

Previous work on identification of ADEs using external sources has largely not been patient-centered. Much of it has focused on discovering adverse reactions from retrospective data. Some have used a single source like the FDA Adverse Event Reporting System (FAERS)[4] data (Tatonetti et al., 2012; Ramesh et al., 2014), while others have worked on information fusion (Sarker and Gonzalez, 2015; Jiang, Solbrig, and Chute, 2011; Yeleswarapu et al., 2014). Their techniques based on large amounts of retrospective data are well-suited for discovering new ADEs, but not applicable for *real-time* ADE detection and attribution.

A few examples do exist for patient-centered ADE-detection in hospital settings. For example, Duke and Friedlin (2010) proposed a real-time decision support service for ADEs, but their system does not handle unstructured data of the type obtained from triage notes. Moreover, these tools do not perform any entity normalization beyond ontology-based mapping to UMLS concepts. As shown in Fig. 3.2, this leads to missing out on crucial evidence of adverse reactions.

3.3 Overview

With the motivation thus presented, we devote this section to presenting an overview of the CDS system for ***i***dentification, ***at***tribution, and ***r***anking ***o***f ***ADE***s (IATRO-ADE). It is designed to push suggestions/recommendations of ADE diagnoses to the physician in a completely non-intrusive manner. Further, it attempts to eliminate – or at least mitigate – alert fatigue (*i.e.*, the clinician becomes less responsive to automated messages over a course of time) by ranking the suggestions by the likelihood of a drug (or multiple drugs) causing the patient's symptoms. Our application, IATRO-ADE, features the simultaneous use of not just multiple knowledge bases (KBs), but multiple

[4] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

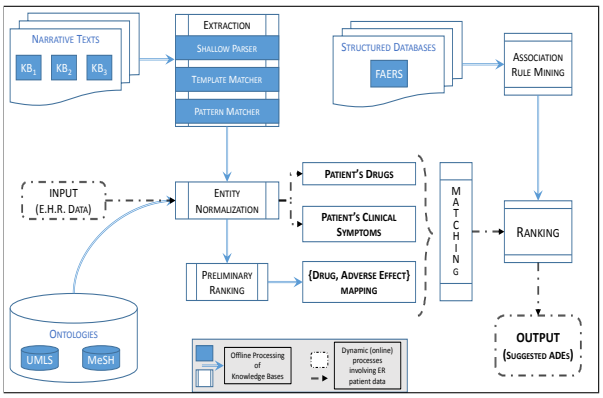

Figure 3.1: Process Flow: The ADE detection system involves (i) offline information extraction from various knowledge bases, and (ii) online processes to handle patient data provided as dynamic input. The final output is a ranked list of ADE attributions provided to the physician.

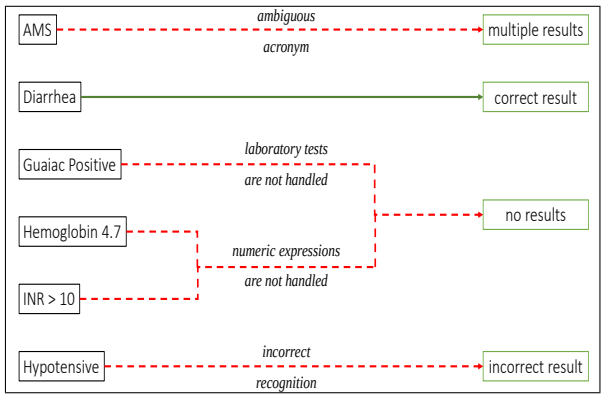

Figure 3.2: Entity Normalization: Current ontology-based information extraction fails to map varied natural language expressions of patients' symptoms to canonical entities. As a result, test results, ambiguous abbreviations, etc. (*e.g.* "Guaiac positive", "AMS") are ignored.

types of KBs to attain this goal. To this end, we make extensive use of natural language processing, template-based mining, and disproportionality analysis metrics.

The processes may broadly be seen as performing the three following tasks:

- $\mathcal{T}_1$: distill the potential adverse effects of a drug from unstructured and semi-structured narratives obtained from multiple KBs,
- $\mathcal{T}_2$: normalize entities using structured ontologies and semi-structured KBs to resolve surface differences of medical entity mentions, and
- $\mathcal{T}_3$: measure the strength of association between drugs and adverse events using structured and unstructured data sources.

In order to find the possible ADE diagnoses, we map the adverse effects (extracted by $\mathcal{T}_1$) of a patient's drugs to her symptoms (processed by $\mathcal{T}_2$). Subsequently, the ADEs are ranked based on the severity of the adverse effect and disproportionality analysis metrics as computed by the third task $\mathcal{T}_3$. The overall process flow is presented in Fig. 3.1.

Entity Normalization

The extraction module processes narrative texts through a combination of shallow parsing and template-based text mining methods. The standard information extraction approach has been to use tools like Metamap (Aronson, 2001) to map matural language to medical concepts in ontologies such as MeSH (Rogers, 1963) or UMLS (Lindberg, Humphreys, and McCray, 1993). This, however, is ill suited for extraction from triage notes because clinicians often use *workplace jargon* that is not captured by existing methods. For example, a clinical symptom may be provided as a laboratory test result like "Hemoglobin 4.7", or a drug name like "simvastatin" may be abbreviated to "simva". Existing methods address the *normalization* of medical expressions, *i.e.* mapping various linguistic expressions to an unambiguous canonical form, in a limited manner. This leads to non-recognition of ADE evidence. Fig. 3.2 provides an example where key evidence of an adverse reaction cannot be correctly detected without extensive entity normalization. Details of the information extraction and entity normalization processes are provided in Sec. 3.4 and Sec. 3.5, respectively.

Offline Processes	Dynamic (Online) Processes
• Extraction from *semi-structured* and *unstructured* data (a) Adverse drug effects (b) Drug-drug interactions (c) Disease characterization in terms of symptoms (d) Laboratory test information • Mining *structured* data (a) Statistical association measures for pairs of the form ⟨`drug, adverse-effect`⟩	• Symptom similarity resolution • Abbreviation resolution and disambiguation • Entity normalization (a) for patient symptoms and complaints (b) for medications • Evidence-based ADE detection and attribution • Ranking ADE diagnosis recommendations

Table 3.1: The two components of the IATRO-ADE pipeline and their constituent processes.

Matching and Relevance Ranking

The complete list of all those drugs that are being taken by a patient and whose adverse effects match the patient's complaints and symptoms is often a large fraction of the patient's entire drug regimen. In other words, simply matching the patient's complaints and symptoms to the side effects of her drugs will almost certainly detect an adverse effect (*i.e.* a very high recall), but is likely to suffer from very low precision. This approach has been shown to cause *alert fatigue* because the CDS system provides too much information of low clinical significance (Beeler, Bates, and Hug, 2014). To resolve this, IATRO-ADE filters out spurious ADE attributions, and the remaining are ranked. Section 3.6 explains this process in greater detail.

3.3.1 Methodology

We use unstructured, semi-structured and structured KBs for the complete pipeline, the key steps of which are divided into an offline component for information extraction from various KBs and a dynamic online component (shown with dashed lines in Fig. 3.1) that handles patient input data. The constituent processes of these two components are listed in Table 3.1. The results obtained in the offline processes are required at various stages by the online component.

IATRO-ADE uses multiple *types* of KBs. The first comprises of drug description repositories. The core knowledge of adverse effects of single drugs and drug-drug interactions (DDIs) is extracted from them. The second type of KBs are medical encyclopedias, which act as knowledge sources for symptoms, diseases and laboratory tests and procedures. Information extracted from these KBs forms the gold standard knowledge for laboratory test results and for characterization of diseases and syndromes in terms of clinical symptoms (*e.g.* "hypotension" is mapped to "dizziness", "fainting", etc.).

Given a set of drugs $D = \{d_1, d_2, \ldots, d_m\}$ and symptoms $S = \{s_1, s_2, \ldots, s_n\}$ obtained from a patient record, IATRO-ADE identifies a subset $D_0 \subseteq D$ such that the adverse effects of drugs in D_0 explain all the symptoms in S. If a symptom $s_j \in S$ can be explained as a possible adverse effect of $d_i \in D$, we denote it as $d_i \xrightarrow{cause} s_j$. For ease of notation, we also extend it to sets of drugs and symptoms so that $D \xrightarrow{cause} S$ denotes that the set of drugs D may cause the symptoms S. It is possible that some symptoms are not known to be adverse effects of any $d \in D$. In such cases, our method attempts to find the subsets of D that can explain maximal subsets of S. Formally, this can be expressed as

$$\mathbb{D} = \{D_i \subseteq D\} = \underset{D' \subseteq D}{\operatorname{argmax}} \left\{ |S'| : S' \subseteq S, D \xrightarrow{cause} S \right\}. \tag{3.1}$$

Note that the $argmax$ is not unique. For example, the patient's entire drug regimen, D, is trivially

in $\mathbb{D}$. Thus, $\mathbb{D}$ is a set of subsets of D, from which we select the minimal sets:

$$D_{min} = \underset{D_i}{\operatorname{argmin}} \{|D_i| : D_i \in \mathbb{D}\}. \tag{3.2}$$

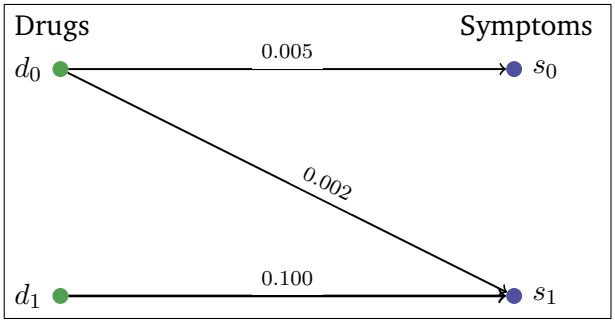

Figure 3.3: The ADE attribution problem: a hypothetical case of a patient taking drugs $\{d_0, d_1\}$ exhibiting symptoms $\{s_0, s_1\}$. The directed edges represent causality between drug and symptom, with edge weights reflecting the likelihood of the cause. Even though d_0 *explains* all symptoms, the partial match $\langle d_1, s_1 \rangle$ is more likely.

Note that eq. 3.2 only attempts to find the minimal subsets with no regard for severity or likelihood of the adverse effects. This naïve approach is clearly not realistic. For instance, consider a patient taking two drugs d_0 and d_1 and complaining of headache and abdominal pain, where both are potential adverse effects of d_0. The drug d_1, however, can only cause the latter. If these adverse effects of d_0 are rare while d_1 causing abdominal pain is common, then it is more likely that the headache is caused by some non-iatrogenic factor and $d_1 \xrightarrow{cause}$ "abdominal pain" is the correct attribution. Fig. 3.3 illustrates this as a weighted bipartite graph.

Data: Set of drugs D and symptoms S
Result: ADE attributions in descending order of coverage of S
$Q_D \leftarrow$ empty queue of sets of drugs;
$\mathcal{P}(D) \leftarrow \{D' \subseteq D\}$;
while $P(D) \neq \emptyset$ **do**
  $D_i = \underset{D_i \in P(D)}{argmax} \left\{|S'| : S' \subseteq S, D' \xrightarrow{cause} S'\right\}$;
  $\mathcal{P}(D) \leftarrow \mathcal{P}(D) \setminus D_i$;
  if $\nexists q \in Q_D, D_i \subseteq q$ **then**
    $Q_D \leftarrow Q_D \cup \{D_i\}$;
return Q_D

Algorithm 3.1: ADE attribution search.

In light of such scenarios, instead of simply solving eq. 3.2, we use it to iteratively compute the list of all possible ADE attributions, as shown in Algorithm 3.1. In order to automatically attribute a patient's symptoms to an ADE, the terms in the information extracted from the drug description repositories are *normalized*. This is the process of mapping varying linguistic expressions to unambiguous canonical forms by similarity resolution, abbreviation resolution and disambiguation, and finally, named entity normalization. For example, if the offline extraction has obtained "fatigue" as the adverse effect of a drug, while an ER patient taking that drug is complaining of "tiredness", similarity resolution identifies the two terms as nearly equivalent. Similarly, expressions such as "anemia", obtained from offline extraction from drug description datasets, are equated to terms like "low hemoglobin" in the patient's health record.

Symptoms expressed as acronyms or abbreviations are resolved by either looking up structured data tables, or by applying the abbreviation extraction algorithm of Schwartz and Hearst (2003) on the third type of KBs used in our application: biomedical literature. For this, our source is the PubMed Central (PMC) repository. Ambiguous abbreviations are resolved (see section 3.5), and the canonical forms are matched to see if a particular drug d_i can cause a symptom s_j.

Finally, a fourth type of KB, structured data from FAERS, is engaged. This is a database that contains information on adverse event and medication error reports submitted volunatrily by clinicians to FDA. It is designed and intended to support the FDA's post-marketing safety surveillance program for drug and other therapeutic products. This dataset, in conjunction with PMC, is used to compute statistical association measures. These measures, in turn, are applied to provide the final ranked list of ADE attribution diagnoses.

3.4 Information Extraction

In the scope of this work, we extract information to map (i) drugs to their adverse effects, (ii) diseases to their symptoms, and (iii) laboratory test results to the diseases and symptoms indicated by abnormal findings. In this section, we present the details of these information extraction processes and their place in IATRO-ADE.

To extract information about clinical drugs, we employed several publicly available online data repositories as well as proprietary services used by the Stony Brook University Hospital. DrugBank[5], consisting of $7,759$ drug entries, was used as the basis of all possible clinical medications. Beyond DrugBank, we also extracted information about these medications from Drugs.com[6], RxList.com[7], Lexicomp and Micromedex (the last two are services used by the SBU hospital). These sources provide a mix of structured and semi-structured information. Structured information was directly extracted using web-scraping techniques. This step was a rule-based approach that relied on the underlying HTML structure of the source pages.

A **template** connecting a drug r to the symptoms s_i of its adverse effects: "(side|adverse) (reactions|effects) .* $\langle r \rangle$.*: $s_1, \ldots, s_{k-1}$ and s_k"
• The following additional **adverse reactions** have been identified during postapproval use of simvastatin: *pruritus*, *alopecia*, *rhabdomyolysis*, ... [Source: RxList.com]
• If any of the following **side effects** occur while taking simvastatin, check with your doctor: *dizziness*, *fainting*, *irregular heartbeats*, ... [Source: Drugs.com]

Table 3.2: The regular-expression of a typical template (top) and two examples from semi-structured knowledge bases. Linguistic cues are in bold, while the extracted symptoms (s_i) are italicized.

Where information was semi-structured, relevant information was available in narrative texts under suitable headings. Examples include "What are the possible side effects of ...?" for information on the possible adverse effects of a drug, or "What are the precautions when taking ...?", for information on drug interactions and contraindications.

In this work, we decided against manually crafting templates since such an approach is clearly not scalable accross multiple KBs for proper information fusion. Instead, we adopted the approach of learning the templates of sentences that contain relevant information. We observed that such sentences usually contained lexico-syntactic patterns indicative of the information being presented. Cues like "may cause" and "side effects" were common. Similarly, syntactic patterns such as long conjunctions of entities of the same semantic type also appeared frequently. These patterns were often very similar to the observations on hyponymy structures made by Hearst (1992), albeit with variations. A simple template that combines such lexical cues with a Hearst pattern, along with two examples of matching text snippets, is presented in Table 3.2.

3.4.1 Learning relation templates from semi-structured information

In traditional information extraction (IE) systems, the local context around words or phrases is used in conjunction with global information (Califf and Mooney (2003), Bunescu and Mooney (2004), Maslennikov and Chua (2007), and Patwardhan and Riloff (2007), among others). Our approach is similar in spirit to the two-stage pipeline process presented by Patwardhan and Riloff (2007). In their work, only those sentences that are identified by a classifier as relevant, are passed on to the IE module. Similarly, we used a three-step pipeline. The first step was to identify the sections in

[5] Knox et al. (2011) Available: http://www.drugbank.ca/
[6] http://www.drugs.com/
[7] http://www.rxlist.com/

the semi-structured KBs that are relevant for a particular kind of information. Then, a classifier is used to identify the relevant sentences from that section. Our final step is the construction of dependency and syntactic parse trees for these sentences, and then identifying the lexico-syntactic constructs that are indicative of the relation of interest.

Note that unlike the body of work cited above, we can obtain the medical entity mentions by using MetaMap (this step is often called "role filling" in standard IE terminology), and thus directly focus on learning the templates that are characteristic of certain relations. In this sense, our template-learning approach resembles the integration of named entities into pattern learning, as done by Filatova, Hatzivassiloglou, and McKeown (2006). Also worth noting is that more recent work on learning templates involves building unified models instead of the pipeline approach (*e.g.* Patwardhan and Riloff (2009) and Chambers and Jurafsky (2011)). In our scenario, however, due to the availability of semi-structured data, the relatively simpler pipeline paradigm suffices. As our evaluations demonstrate later, we use it successfully to complete the three types of mappings underscored earlier, *viz.*, map (i) drugs to their adverse effects, (ii) diseases to their symptoms, and (iii) laboratory test results to the diseases and symptoms indicated by abnormal findings. In the remainder of this section, we describe the three steps of the IE pipeline introduced above.

As mentioned before, these KBs are *semi-structured* in the sense that each entry in a KB is divided into labeled sections, where each section presents a certain type of information. In the drug data repositories, for instance, sections had labels such as "side effects", "interactions", "what are the precautions when taking . . .", etc. These labels are consistently used within each KB, so a small number of fixed rules were sufficient for identification of the relevant sections.

Within each section, however, the data was presented in natural language text. To identify the sentences that actually express the relations pertaining to our application, we built a linear kernel support vector machine (SVM) classifier using LIBLINEAR (Fan et al., 2008). To this end, we performed lemmatization, and then used unigrams, bigrams and trigrams as lexical features. Additionally, phrases identified by MetaMap as medical entities were also included. Now, because the feature vectors were being built for sentences that were already known to be from a section discussing only one particular kind of relation (*e.g.* adverse drug effects), we were able to exploit the information obtained from the structured portions of these datasets and assume the distant supervision hypothesis: if a sentence contained an entity-pair known to be in a certain relation, we assumed that the sentence was actually expressing that relation. Thus, using a small number of relations extracted from structured data, we were able to bootstrap a classifier that labeled each sentence as (ir-)relevant with respect to a particular relation.

The final step of this IE pipeline is to extract entity-pairs from the relevant sentences. Again, a linear kernel SVM classifier was built for this. But instead of labeling sentences, this was used to label entity-pairs. To illustrate this, consider the extraction of (drug, adverse-effect) pairs from a sentence. If a pair was already among those obtained from structured data, it was marked as true. Then, the sentence was parsed and the dependency path connecting the drug to its marked adverse effect was extracted. Finally, the medical entities were replaced by their semantic types. These paths formed the training data. Once the trained model was available, all the sentences deemed relevant in the previous step were passed through MetaMap, parsed, and the dependency paths obtained from these parse trees were provided as input to this classifier. A simple example of extracting such a path is shown in Fig. 3.4.

This IE pipeline was used to map drugs to their adverse effects and drug-pairs to their interactions. Some KBs like Drugs.com and Micromedex also provide a 'severity' calue for drug interactions. This, too, was extracted for later use in our ranking algorithm.

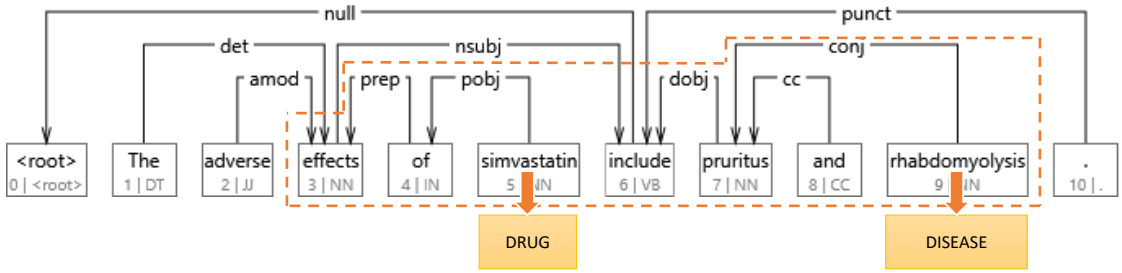

Figure 3.4: The entity-pair ⟨simvastatin, rhabdomyolysis⟩ was obtained from structured data as a gold-standard (drug, adverse-effect) instance. Thus, in sentences like "The adverse effects of simvastatin include pruritus and rhabdomyolysis.", the dependency path joining them (dashed portion of the parse tree) was used to train our template-learning classifier after the entity instances were replaced by their semantic types ("drug" and "disease", respectively).

Characterizing diseases in terms of their symptoms

As we saw in Table 3.2, an adverse effect of a drug may be given in terms of symptoms (*e.g.* "dizziness") or diseases/disorders (*e.g.* "rhabdomyolysis"). Similarly, the patient data may also have symptoms as well as diseases. Thus, a crucial piece of the IATRO-ADE pipeline is to understand the manifestation of a disease in terms of its clinical symptoms. Continuing with the previous example (Table 3.2 and Fig. 3.4), let us consider an ER patient on simvastatin and complaining of "muscle pain", "joint pain" and "tiredness". All three are symptoms of "rhabdomyolysis", an adverse effect of the drug. But since the ADE information extracted so far does not map "rhabdomyolysis" to the observable symptoms, this will not be detected.

The technique of extracting the symptom-based characterization of a disease or disorder is identical to the template learning process described for mapping drugs to their adverse effects. The KBs, however, are general-purpose medical encyclopedias instead of drug databases: MedicineNet[8] and MedlinePlus[9].

Mapping laboratory test results to diseases and symptoms

Here, we describe the last leg of the three-pronged approach: mapping laboratory test results to the diseases and symptoms corresponding to their abnormal findings. It is important to note that even though we usually think of symptoms as observable physical or mental states like seizures, headaches, etc., a wide range of symptoms are expressed in terms of medical tests (Banerjee et al., 2014). An abnormal test result is often an indication or confirmation of a symptom or disease. In the remainder of this work, we use the term 'test' to mean laboratory tests as well as comparatively simple readings such as pulse rate, blood pressure, etc. As illustrated earlier in Fig. 3.2, hospital records often use such evaluations to express the presence of a medical condition. For example, the symptom "hemoglobin 4.7" indicates a very low red blood cell count. In order to infer this, however, IATRO-ADE needs to know the reference range for hemoglobin. This section explains how we extract such information and draw meaningful inferences from the results.

We use three KBs to map test results to symptoms: (i) the list of procedures and tests available on MedicineNet, (ii) the Laboratory Test Database at University of California, San Fransisco[10], and

[8] www.medicinenet.com/

[9] https://www.nlm.nih.gov/medlineplus/

[10] UCSF Departments of Pathology and Laboratory Medicine | SFGH Lab Manual | Laboratory Test Database. Available:

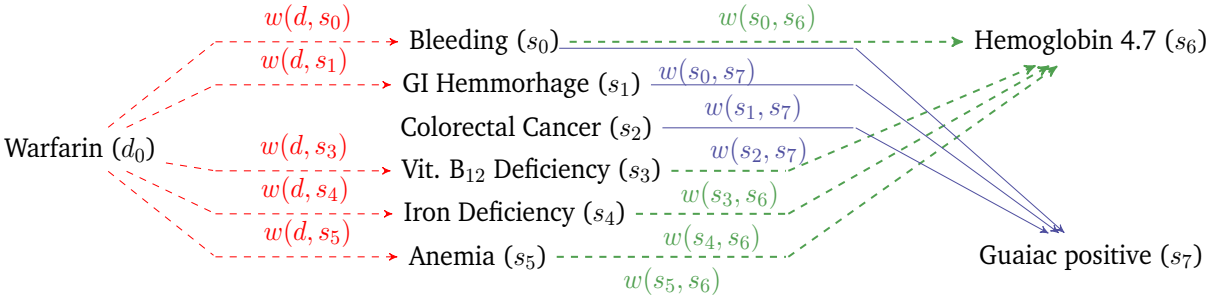

Figure 3.5: A graphical model of ADE attribution for a single drug, Warfarin (d_0), based on two laboratory test results. Edge direction denotes causality, and the likelihood of a cause-effect relation between nodes m and n is encoded as edge weight functions $w(m, n)$. Note that multiple test results are capable of reinforcing ADE signals. In this example, most symptoms indicated by the hemoglobin count (s_6) and guaiac positive value (s_7) point to an adverse effect of Warfarin. In case of uniform likelihoods, the most probable adverse effect is "bleeding" (s_0).

(iii) the health encyclopedia available from University of Rochester's Health Encyclopedia[11].

Compared to the extraction of adverse effects, test results are harder to interpret. This is because reference ranges vary depending on the patient's age, gender and medication history. They may also vary from one laboratory to another. Our work errs on the side of caution, and if even one of the KBs claims a particular value to be abnormal, we consider it to be so. In other words, if different KBs disagree on a reference range, we take the most conservative estimate.

The extraction process comprises of obtaining noun phrases using the Genia tagger/parse (Tsuruoka et al., 2005), and then using template-rules to selectively pass these phrases on to be labeled by MetaMap. This leads to the discovery of two kinds of knowledge:

(a) the reference range or value (*i.e.*, the normal range/value of the measurement), and
(b) what abnormal values may indicate

For example, from "hemoglobin 4.7" in a hospital record, IATRO-ADE is able to infer that (a) patient's hemoglobin count is *lower* than the normal value, and (b) the possible reasons are (to name a few) *anemia*, *bleeding*, *stomach ulcer*, or *iron deficiency*. Moreoever, in cases where multiple results point to a common reason, the ability to automatically infer these results enable our approach to distill stronger signals from all available data, and thus identify the most likely ADE. Fig. 3.5 illustrates how these likelihoods are computed.

3.5 Entity Normalization

So far, we have described how we extract different types of information pertaining to diseases, symptoms, laboratory tests and adverse effects of drugs – using multiple information sources for each. Even though fusing such heterogeneous information ensures significantly higher coverage, it leads to extensive polysemy, where vastly different linguistic expressions may refer to the same medical concept. In some cases, resolving the surface differences is simply an issue of identifying and linking synonymous (*e.g.*, the generic name "warfarin" and its brand name "Coumadin") or nearly-synonymous (*e.g.*, the symptoms "breathlessness" and "shortness of breath") entities. Frequently, less obvious equivalences need to resolved, however. These arise from (a) the pervasive use of domain-specific abbreviations, and (b) expressions involving numeric (*e.g.* "hemoglobin 4.7", "INR 7.1") or nominal (*e.g.* +/-, high/low) values.

http://labmed.ucsf.edu/sfghlab

[11] http://health.rush.edu/HealthInformation. Accessed: Nov 14, 2013

Abbreviation	Expansion	UMLS concept
NCRS	Nutrition-related chronic diseases	–
CB1	Cannabinoid-1	CNR1 gene
SGLT2	sodium glucose co-transport-2	SLC5A2 gene
vit. def.	Vitamin Deficiency	VIT gene; butyl phosphorotrithioate
GIB	gastrointestinal bleeding	(Gibraltar) [Geographic Area]

Table 3.4: Limitations of ontology-based entity normalization: many abbreviations in our data were either absent from the UMLS meta-thesaurus, or were mapped to incorrect concepts by MetaMap.

3.5.1 Linking synonyms and near-synonyms

This is by far the simplest of our entity normalization processes due to the significant amount of prior research devoted to building general linguistic ontoligies like WordNet (Miller, 1995) as well as biomedical knowledge ontologies like SNOMED-CT, NDF-RT, RxNorm and MeSH. Since the Unified Medical Language System (UMLS) integrates over a 100 such medical ontologies, we exclusively work with UMLS for identifying synonymous medical terms. For identifying equivalent or nearly-equivalent terms that are less domain-specific, we also engage WordNet.

The UMLS meta-thesaurus maps each medical term to a *unique concept identifier*, a unique *preferred* name and one or more *semantic types*, which are biomedical categories. It holds over 12 million concept names collated into 3.1 million unique concepts and categorized into 135 semantic types.[12] Exploiting this massive repository allows us to identify most variations in medication and disease names. To a large extent, it is also able to link synonymous symptoms like "shortness of breath" and "dyspnea". Table 3.3 presents the recall of UMLS-based entity identification on our data, showing nearly perfect identification of all drugs except when non-standard abbreviations are used, *e.g. simva* and *amio* for *simvastatin* and *amiodarone*, respectively. The ability to correctly identify patients' symptoms and complaints, however, suffers more. This is because a large fraction of symptoms are expressed in terms of laboratory test results, which the existing medical ontologies do not map to symptoms.

Entity Type	Recall	Errors	Cause
Drugs	0.98	*Amio* *Simva* *VPA*	non-standard abbreviations
Symptoms	0.69	*qtc* > 460 *occult stool*	no knowledge of laboratory test data

Table 3.3: Named entity normalization of terms in patient health record, based on the UMLS meta-thesaurus. Non-standard use of abbreviations and extensive use of laboratory test results lead to imperfect (and for some semantic types, poor) recall.

3.5.2 Abbreviation Resolution

A fairly comprehensive list of abbreviations is already present in UMLS, but as shown in Table 3.3, a better entity normalization process requires intelligent abbreviation resolution that goes beyond simply looking up ontologies. The first step in this direction is to check whether an abbreviation matches a synonymous term instead of just the list of patient's medications. This is done to normalize names such as 'VPA' and 'divalproex'.

Non-standard abbreviations, however, cannot be resolved in this manner. We thus implement the algorithm proposed by Schwartz and Hearst (2003) to identify abbreviation definitions from

[12] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

unstructured medical texts. This algorithm is run on the PubMed Central (PMC) dataset, comprising of more than 2.7 million research articles. Table 3.4 shows a few abbreviations we found by this method that were either absent in UMLS or were incorrectly labeled by Metamap. It is often the case that a single abbreviation has multiple possible expansions. The observation made in such cases was that even though all expansions were sensible in their own right, there was, in each case, only one expansion that was most relevant to the patient's medications or symptoms. This was a clear indication that in order to identify the correct expansion, and subsequently, normalize entities, it was imperative that the *context* of the abbreviation be taken into account. To learn the correct expansion in a context-dependent manner, we built a distributional semantic model using the PMC dataset and the semi-structured datasets described in Section 3.4.

Distributional Semantic Modeling

Distributional semantic models rely on the *distributional hypothesis*, which states that *the meaning of a word is the set of contexts in which it occurs across a large number of texts*. To illustrate this clearly, we present two examples in Table 3.5, one medical and one non-medical, to show how the meaning of a previously unknkown word may be inferred from its context. This approach has been used in non-medical recommender systems before (Herlocker and Konstan, 2001; Adomavicius and Tuzhilin, 2011) and very recently, also been applied to the entity normalization problem (Musto et al., 2014; Lipczak, Koushkestani, and Milios, 2014).

(I)	After 10mg of ___, *hyperglycemic patients* felt immediate *relief*. Metformin is known to provide *relief* to *patients* suffering from *hyperglycemia*.
(II)	The *tiny brown* ___ *scurried away under the branches*. The *little brown* rabbit *disappeared under the branches*.

Table 3.5: The distributional hypothesis: (I) suggests a treatment that lower blood glucose levels, and (II) implies that the subject is most likely a small animal.

Given an abbreviation, our goal is to identify the correct expansion out of multiple candidate expansions. In order to achieve this, we construct a context vector v_e for entity e. A *context vector* of a term is a vector designed to capture the set of contexts in which it occurs. For IATRO-ADE, we build these vectors by considering a fixed-size window around every mention of e. For documents in the PMC dataset, the paragraph in which e occurs is considered as the context window, and for documents from the semi-structured sources, the text contained under the relevant section heading (*e.g.* "adverse effects") is considered. All context vectors are normalized to unit length, and the angular distance based on cosine similarity is used to determine the distance between two vectors:

$$d(v_e, v_f) = 1 - cos^{-1}\left(\frac{v_e.v_f}{||v_e||.||v_f||}\right)\frac{1}{\pi} \tag{3.3}$$

where $||x||$ denotes the Euclidean norm of a vector. The candidate expansion whose context vector is closest to the context vector of the abbreviation is selected.

It should be noted that creating the context vectors is a resource-intensive process, and cannot be achieved in real-time scenarios. Context vectors of abbreviations in the KBs are thus constructed offline, and all the candidate expansions are retained. The final selection is done based on the context provided by patient data, *i.e.*the list of medications and symptoms. For a given abbreviation a, its expansion e_a can thus be formally expressed as

$$e_a = \underset{e \in D \cup S}{\operatorname{argmin}}\, d(v_e, v_a) \tag{3.4}$$

where D and S are the sets of drugs and symptoms obtained from the patient record.

3.5.3 Normalizing expressions involving laboratory test results

As we saw before, a patient's symptoms may be expressed in terms of laboratory test results. This means that phrases like "hemoglobin 4.7", "guaiac positive" (Fig. 3.2) or "occult stool" (Table 3.3) must be identified with the canonical name of the corresponding condition. Just like in abbreviation resolution, this mapping, too, generates multiple candidates.

Unlike abbreviation resolution, however, we do not select one candidate as the canonical expression. Instead, all the possible causes extracted from semi-structured KBs (using methods described in Sec. 3.4) are retained, and a probabilistic identification is induced. For example, if an abnormal test result indicates n distinct potential conditions (*e.g.* low hemoglobin may indicate gastrointestinal bleeding, iron deficiency, etc.), a probability is associated to each potential cause. The probability distribution is computed by extracting detailed syntactic and semantic information from the PMC dataset, and subsequently calculating co-occurrence statistics.

For each symptom expressed in terms of a laboratory test result, the lexical content is separated from the expression. For example, from "hemoglobin 4.7", we retain "hemoglobin". Further, ubiquitous terms like "test", etc. are also removed in order to retain only the terms specific to the symptom. For example, if a symptom is expressed as "guaiac test was positive", we will only retain the term "guaiac". The filtering is done by tokenizing all the text from the laboratory test KBs, and ranking all terms by their *inverse document frequency* (IDF). For D being the set of all documents under consideration, the IDF of a term t is defined as

$$IDF(t) = \log\left(|D|/|\{d \in D : t \in d\}|\right). \tag{3.5}$$

This measures whether t is common or rare across all documents. Common terms, clearly, are not significant for a specific test. We can thus filter out generic words like "test", "was", etc. Letting t_0 denote the term that remains after the above filtering is carried out, we search the PMC dataset to find (i) the number of documents in which t_0 appears, and (ii) the number of documents in which a potential cause (*e.g.* iron deficiency) co-occurs with t_0.

In case the original symptom was a numeric expression, or contained an adjective modifier, we marked the semantic orientation of this expression. This was done for nominal expressions (*e.g.* +/-) by triggering binary values. For numeric expressions, we checked whether the value was lower or higher than the reference range, and trigger a binary value indicating *low/high*. For example, for the expression "hemoglobin 4.7", we compare the value 4.7 with the reference range for hemoglobin test, and mark the fact that 4.7 is lower than normal. Let this semantic orientation be denoted by $S(t_0)$. When querying the KB, we filtered out documents unless the same semantic orientation $S(t_0)$ was observed for t_0 in the document.

This was achieved by splitting the text into sentences, and then generating the dependency parse tree of those sentences in which t_0 appeared. We then checked if a word with a negative connotation, like "low", was a modifier of t_0. This check was performed after lemmatization so that inflected forms like "lower", "lowered", etc. were accounted for. In case there of a numeric dependency, the semantic orientation was checked by comparing it to the reference range for the test. Let the semantic orientation obtained from the dependency parse tree be denoted by $T(t_0)$. The probability of a cause c is then given by

$$P(c|t_0) = \frac{|\{d \in D : t_0 \in d,\, T(t_0) = S(t_0),\, c \in d\}|}{|\{d \in D : t_0 \in d,\, T(t_0) = S(t_0)\}|} \tag{3.6}$$

These probability values were used as illustrated earlier in Fig. 3.5. They were also used to rank the patient's medications by their relevance with respect to the reported symptoms. This ranking process is described in details in the next section.

3.6 Ranking

The ADE attributions are computed using the exhaustive search Algorithm 3.1. The extensive steps taken to extract relevant information from several KBs and subsequently normalize entity names is necessary for the matching algorithm to perform well, and exhibit high recall (*i.e.* it does not miss a potential ADE attribution). However, as prior work in this domain has shown, presenting the physician with all such potential cases leads to alert fatigue. IATRO-ADE thus employs a ranking algorithm as well. In this section, we discuss this, and show that the most plausible attributions are always ranked among the top diagnostic recommendations provided by IATRO-ADE.

The first step is to rank the possible ADEs by severity. The severity ratings were extracted from semi-structured KBs as discussed earlier in Section 3.4. The next step is to obtain co-occurrence statistics for drugs and their adverse effects. To this end, we made use of the structured portions of the semi-structured KBs, where relevant statistics are often presented in tabular format. For example, for several drugs, the percentage of reported cases where patients suffered from a particular adverse effect are reported in this manner.

To bolster the association between drugs and their adverse effects, IATRO-ADE also performs simple disproportionality analysis (DPA) on the FAERS dataset. The first DPA metric we used is a conditional probability value known as the *relative reporting ratio* (RRR). It computes the ratio between the probability of a symptom s given a drug d and the probability of s in the entire dataset:

$$RRR(s,d) = \frac{P(s|d)}{P(s)} = \frac{P(s,d)}{P(s).P(d)}. \tag{3.7}$$

The second is a relative risk metric, called the *proportional reporting ratio* (PRR). It measures the ratio of the frequency of s in patients exposed to d to the frequency of s in unexposed patients:

$$PRR(s,d) = \frac{P(s|d)}{P(s|\neg d)}. \tag{3.8}$$

Both metrics were computed only for those $\langle s,d\rangle$ pairs that occurred at least 5 times, as co-occurrences that are too rare are not deemed statistically meaningful. These frequency-based association scores were also computed on the PMC dataset. However, our initial experiments showed that if, while computing the joint probability score $P(s,d)$, we insist that s and d are syntactically connected, then the RRR and PRR scores tend to be negligible. Further, we noticed that a much stronger signal is obtained if the condition is relaxed to include every document where s and d co-occur. This relaxed notion of co-occurrence was used to compute the RRR and PRR metrics based on the PMC dataset. In addition to computing these two metrics to capture drug-symptom or drug-disease associations, we also used them to compute association metrics for drug-drug interactions. This was done with a simple extension of the formulae in eq. 3.7 and 3.8. Thus, the IATRO-ADE ranking scheme stands on three measures:

(m_1) severity
(m_2) RRR and PRR based on FAERS and tabular data from semi-structured KBs
(m_3) RRR and PRR based on PMC data

The likelihood $\mathcal{L}(d,s)$ of a drug d causing a symptom s is a weighted linear sum of these three factors whose coefficients were experimentally determined:

$$\mathcal{L}(d,s) = \alpha m_1 + \beta m_2 + \gamma m_3. \tag{3.9}$$

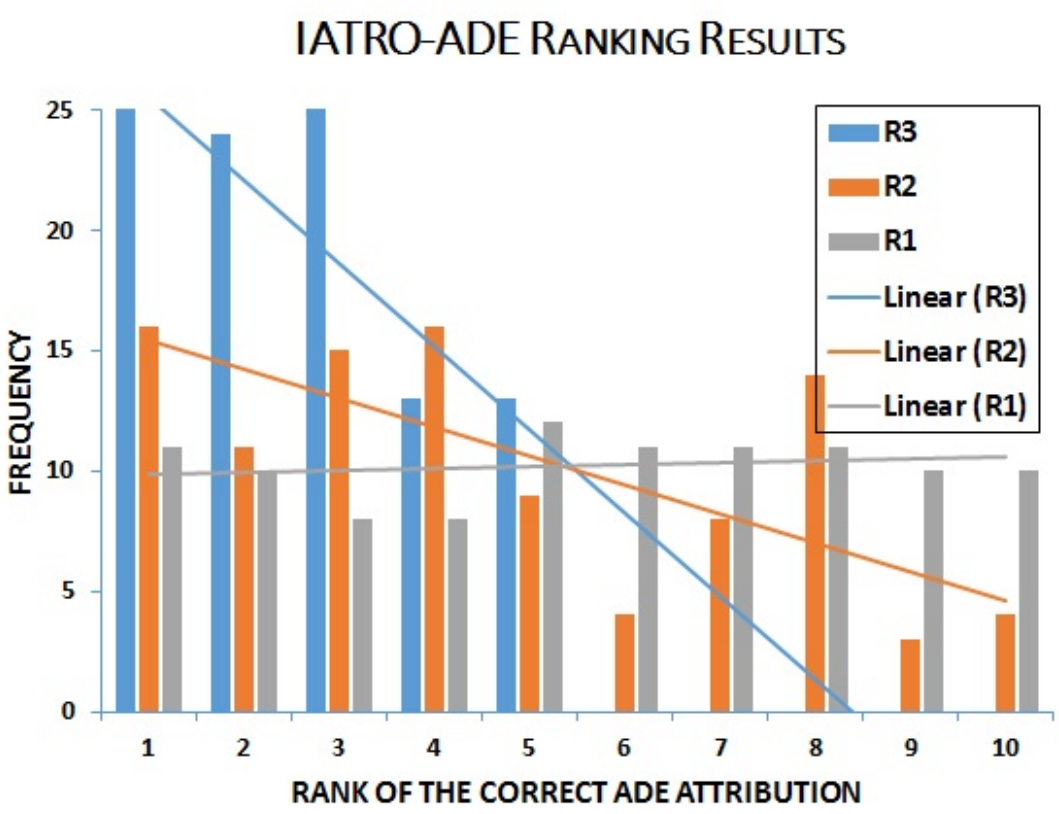

Figure 3.6: Distribution of the ranks of the correct ADE attributions computed by IATRO-ADE. Shown here are R_1 (only severity rating, *i.e.*, $\beta = \gamma = 0$ in eq. 3.9), R_2 (severity plus DPA on structured and semi-structured KBs, *i.e.*, $\gamma = 0$), and R_3 (severity plus DPA on all types of KBs, *i.e.*, $\alpha, \beta, \gamma > 0$). The linear trends show that R_3 is successful in providing the correct ADE attributions at the top of the ranked list. In our data, IATRO-ADE is always able to provide the correct ADE attribution among the top 5 suggestions.

3.7 Experimental Results

We now present the experimental results of IATRO-ADE. The evaluation was done by human experts on a dataset of 100 ER patient records. As a baseline, we performed maximal matching (MM), where the output consists of all possible drugs that can cause the given symptoms. No entity normalization or ranking is done for the baseline, providing a naïve and imprecise system that returns every potential ADE that can be directly matched to the patient's drug(s).

Note that even though the baseline yields every potential adverse effect, it still suffers from low recall because a large fraction of drugs and conditions cannot be identified using just medical ontologies. The missed out cases include laboratory test results, non-standard abbreviations and phrasal expressions. To see how these expressions can also be identified, we continued with maximal matching, but added entity normalization. This method, denoted MM_{EN}, achieves nearly perfect recall. But since it matches everything without performing any filtration, it suffers from very low precision. We then add the iterative ADE attribution algorithm (Algorithm 3.1), denoted IA, to obtain the smallest set of drugs that can explain maximal sets of the patient's symptoms. Finally, we add entity normalization to the IA method, denoted by IA_{EN}.

The ranking algorithm is evaluated separately since precision and recall are set-based measures, computed over unordered collections. The precision and recall values over the entire list of ranked results will be identical to the unranked set obtained by IA_{EN}. To evaluate the ranking algorithm, we use *mean reciprocal rank* (MRR), a widely used metric in the information retrieval domain (Voorhees et al., 1999; Baeza-Yates, Ribeiro-Neto, et al., 1999). It is the average of the reciprocal ranks of all the output lists. If Q is the set of data being evaluated (in our evaluation, $|Q| = 100$), and r_i is the rank of the correct suggestion for the i^{th} patient, then

Experiment	Precision	Recall	MRR		
Baseline	0.23	0.52	–		
MM_{EN}	0.39	0.94	–		
IA	0.65	0.52	–		
IA_{EN}	0.75	0.94	–		
$\text{IA}_{EN,Rank}$	0.75	0.94	R_1 0.30	R_2 0.37	R_3 0.59

Table 3.6: Experimental Results: (i) the precision and recall of IATRO-ADE in different experimental settings, and (ii) the mean reciprocal rank of the correct diagnostic recommendation based only the severity-based ranking (R_1), severity plus disproportionality measures from semi-structured KBs (R_2) and finally, adding disproportionality measures from all types of KBs (R_3).

$$MRR = \frac{1}{|Q|}\sum_{i}^{|Q|}\frac{1}{r_i}. \tag{3.10}$$

The results of our experiments are shown in Table 3.6 using (R_1) only the severity-based ranking, (R_2) severity plus DPA metrics from semi-structured and structured KBs, and (R_3) severity plus DPA metrics from all types of KBs. Additionally, Fig. 3.6 illustrates the distribution of the rank of the correct diagnostic recommendation made by IATRO-ADE. The best results are achieved when severity-based ranking is combined with the disproportionality metrics computed on all the semi-structured KBs *and* the unstructured data from the PMC dataset. Especially, we note that in our data, IATRO-ADE was able to provide the correct ADE attribution for *each* patient as one of the top 5 recommendations to the clinician.

3.8 Summary

In this chapter, we presented a pipeline IATRO-ADE to automatically identify adverse drug events based on a patient's drug regimen and the symptoms and complaints s/he presents. Further, we attribute these adverse events to these symptoms and complaints, and rank the possible diagnoses with respect to their relevance to the individual patient's condition. This process is accomplished by extracting information from multiple *types* of knowledge sources. It involves normalizing medical entities, resolving ambiguous acronyms and abbreviations, and incorporating laboratory test results. Unlike most state-of-the-art clinical decision support systems, our application presents a combination of filtering and ranking in order to avoid suggesting diagnoses of low clinical significance.

We demonstrated the performance of our method on real (de-identified) patient data obtained from the emergency room of the Stony Brook University Hospital, and showed that in each case, the correct diagnosis was always among the top five suggestions. In conclusion, we would like to say that IATRO-ADE is capable of alerting clinicians with highly accurate ADE attribution notifications, and has shown promising results thus far in a clinical setting.

Chapter 4

Conclusion

In this proposal, we presented our contributions toward connecting the dots in what has often been called the *translational* pipeline in biomedicine: the "bench to bedside" journey made by newly discovered knowledge from research settings into actual patient care. It has often been claimed that in spite of substantial research in various areas of biomedical research, there remains a significant *translational gap*, which has led to tragic loss of life and resources on many occasions. So much so that medical errors have caused more fatalities than cancer or heart diseases! Our work presents novel research spanning – and more importantly, bridging – this gap.

The first part of this proposal presented a novel relation inference mechanism that combines the strengths of natural language processing with intuitions from the domain of pharmacological science to learn new drug-disease relations from biomedical research literature, thereby contributing to the first stage of the translational pipeline: the proverbial "bench". In spite of the linguistic complexity of research literature, our approach – the *latent pathway model* – is capable of using extremely simple features to distill relations between medical entities even in the absence of discourse connectives, and even when the related entities (*viz.*, drugs and medical conditions) never co-occur in the same sentence or document.

Professionals in the biomedical domain, and healthcare practitioners in particular, have long deliberated over how the current trend in translational research focuses largely on carrying initial research findings into the clinical investigation stage, but not so much on the later stages of the journey, where the results are actually implemented in patient care settings. With that in mind, the second and third chapters of this proposal focused on the later stages of translational research. There, we described our contribution towards using the knowledge about relations among medical entities to aid healthcare activities in real clinical settings. In particular, our techniques of harnessing diverse biomedical corpora to distill actionable information was used to build clinical decision support systems aimed at improving two highly error-prone areas of diagnostics: the proper and timely incorporation of laboratory tests, and identification and attribution of adverse drug events. With evaluations done on de-identified patient data obtained from emergency room scenarios, we demonstrated that combining the diverse knowledge that is available in various forms and in various stages of the translation pipeline, can have a significant positive impact on patient care.

Relation Inference in Biomedical Literature

This first component of our proposal focused on extracting new relational information directly from the vast body of research literature in biomedicine. We presented a novel global inference framework that *infers* relations between drugs and medical conditions based on the underlying implicit pharmacologic effects of drugs. This model, based on the *latent pathway* of drugs, was designed such that relations could be inferred even when (i) the entities were not in the same sentence or document, and (ii) there was no explicit discourse connecting them to each other. The empirical findings of this work showed that this methodology is consistently capable of inferring relations across multiple drugs and diseases, and yields a high number of new relations without compromising precision. It is a substantial improvement over a sentence-level supervised classification built on the same underlying feature space. Finally, as a *human-in-the-loop* process, the latent pathway model is a quick and viable option for augmenting existing medical data repositories with new knowledge.

Healthcare Applications: the proverbial "bedside"

The next part of our proposal presented two healthcare applications that employ the kind of relational information we strive to learn from research literature. Here, we discussed our contribution towards improving two highly error-prone areas of the last stage of the translational pipeline: incorporating laboratory tests in the diagnostic process, and detecting adverse drug events.

Our key observation here was that even though there are multiple data repositories available for various purposes, the information that is needed for fast and accurate clinical decision support in these areas, is scattered across several different kinds of knowledge bases. Based on this observation, we built two applications wherein heterogeneous data repositories are brought together to incorporate information extracted from structured, semi-structured and unstructured data.

The first healthcare application presented, in chapter 2, is a diagnostic test recommendation system that uses an application-specific entity recognition and a template-based information extraction approach to suggest laboratory tests that can be used to confirm (or invalidate) an adverse drug effect. Our evaluations showed that harnessing a diverse group of knowledge bases and, subsequently, carefully engineering the information extractions processes, yields a system capable of making highly accurate recommendations.

The second application, called IATRO-ADE, is designed to identify adverse drug events based on a patient's drug regimen, and then attribute (if applicable) the patient's symptoms to the adverse effects of one or more drugs s/he is taking. Our work here, again, involved a careful fusion of relevant information from a diverse group of knowledge bases. The process of identifying and attributing adverse drug events comprised of a domain-specific entity normalization step, automatically learning relation templates, and using disproportionality analysis metrics to rank the diagnostic suggestions. We demonstrated, on real patient data, that IATRO-ADE is capable of delivering the correct ADE diagnosis among the top five suggestions in every single case, with the top diagnosis being correct in 25% of the cases, and one of the top three being correct in 74% of the cases.

To summarize, in this proposal, we contributed to the initial and final stages of the translational pipeline by proposing a novel relation inference model based on latent pharamacologic effects of drugs, and then presenting two highly accurate healthcare applications that harness such relations from multiple diverse knowledge bases to improve patient safety.

Bibliography

Adomavicius, Gediminas and Alexander Tuzhilin (2011). "Context-aware recommender systems". In: *Recommender systems handbook*. Springer, pp. 217–253 (p. 48).

Agbabiaka, T., J. Savovic, and E. Ernst (2008). "Methods for causality assessment of adverse drug reactions: a systematic review". In: *Drug Saf* 31.1, pp. 21 –37 (p. 35).

Agichtein, Eugene and Luis Gravano (2000). "Snowball: Extracting relations from large plain-text collections". In: *Proceedings of the Fifth ACM conference on Digital libraries*. ACM, pp. 85–94 (pp. 3, 5).

Airola, Antti et al. (2008). "A graph kernel for protein-protein interaction extraction". In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, pp. 1–9 (p. 3).

Airoldi, Edoardo M et al. (2006). "Mixed membership stochastic block models for relational data with application to protein-protein interactions". In: *Proceedings of the International Biometrics Society Annual Meeting*, pp. 1–34 (p. 8).

Ananiadou, Sophia and John Mcnaught (2005). "Text Mining for Biology And Biomedicine". In: (p. 1).

Andronis, Christos et al. (2011). "Literature mining, ontologies and information visualization for drug repurposing". In: *Briefings in bioinformatics* 12.4, pp. 357–368 (pp. vii, 2).

Aronson, Alan R (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 17 (pp. 13, 25, 40).

— (2006). "MetaMap: Mapping text to the UMLS Metathesaurus". In: *Bethesda, MD: NLM, NIH, DHHS*, pp. 1–26 (p. 28).

Bach, Nguyen and Sameer Badaskar (2007). "A review of relation extraction". In: *Literature review for Language and Statistics II* (p. 5).

Baeza-Yates, Ricardo, Berthier Ribeiro-Neto, et al. (1999). *Modern information retrieval*. Vol. 463. ACM press New York (pp. 33, 52).

Banerjee, Ritwik et al. (2014). "Automated Suggestion of Laboratory Tests for Identifying Likelihood of Adverse Drug Events". In: *IEEE International Conference on Healthcare Informatics*. Institute of Electrical and Electronics Engineers (IEEE) (pp. viii, 45).

Banerjee, Ritwik et al. (2015). "Patient Centered Identification, Attribution and Ranking of Adverse Drug Events". In: *IEEE International Conference on Healthcare Informatics*. Institute of Electrical and Electronics Engineers (IEEE) (p. viii).

Baseman, J. G. et al. (2013). "Public health communications and alert fatigue". In: *BMC Health Serv Res* 13, p. 295 (p. 39).

Batista, David S, Bruno Martins, and Mário J Silva (2015). "Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantics". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal*, pp. 499–504 (p. 5).

Beeler, P. E., D. W. Bates, and B. L. Hug (2014). "Clinical decision support systems". In: *Swiss Med Wkly* 144, w14073 (p. 41).

Berant, Jonathan, Ido Dagan, and Jacob Goldberger (2012). "Learning entailment relations by global graph structure optimization". In: *Computational Linguistics* 38.1, pp. 73–111 (p. 8).

Beuscart, Régis, Werner Hackl, and Christian Nøhr (2009). *Detection and prevention of adverse drug events: information technologies and human factors*. Vol. 148. IOS Press (p. 34).

Beuscart, Régis, Peter McNair, Jytte Brender, et al. (2009). "Patient safety through intelligent procedures in medication: the PSIP project". In: *Stud Health Technol Inform* 148, pp. 6–13 (p. 35).

Birmingham, Karen (2002). "What is translational research?" In: *Nature medicine* 8.7, pp. 647–647 (p. 1).

Boissier, Marie-Christophe (2013). "Benchmarking biomedical publications worldwide". In: *Rheumatology* 52.9, pp. 1545–1546 (p. vii).

Bollacker, Kurt et al. (2008). "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge". In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. New York, NY, USA: ACM, pp. 1247–1250 (p. 5).

Bordes, Antoine et al. (2013). "Translating embeddings for modeling multi-relational data". In: *Advances in Neural Information Processing Systems*, pp. 2787–2795 (p. 8).

Botsis, Taxiarchis et al. (2011). "Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection". In: *J. Am. Med. Inform. Assoc.* 18, pp. 631–638 (pp. 35, 38).

Brin, Sergey (1999). "Extracting patterns and relations from the world wide web". In: *The World Wide Web and Databases*. Springer, pp. 172–183 (p. 5).

Brunton, L., J. Lazo, and K. Parker (2005). *Goodman & Gilman's The Pharmacological Basis of Therapeutics, Eleventh Edition*. McGraw Hill professional. McGraw-Hill Education (p. 11).

Bunescu, Razvan and Raymond J Mooney (2004). "Collective information extraction with relational Markov networks". In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 438 (p. 43).

Bunescu, Razvan C and Raymond J Mooney (2005). "A shortest path dependency kernel for relation extraction". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 724–731 (pp. 3, 4).

Burnand, Bernard (2015). "Improving the implementation of evidence-based knowledge in healthcare". In: *Clinical Investigation* 5.1, pp. 5–7 (p. viii).

Butler, Declan (2008). "Translational research: crossing the valley of death". In: *Nature News* 453.7197, pp. 840–842 (p. 1).

Califf, Mary Elaine and Raymond J Mooney (2003). "Bottom-up relational learning of pattern matching rules for information extraction". In: *The Journal of Machine Learning Research* 4, pp. 177–210 (p. 43).

Cantor, Michael N., Henry J. Feldman, and Marc M. Triola (2007). "Using trigger phrases to detect adverse drug reactions in ambulatory care notes". In: *Qual. Saf. Health Care* 16.2, pp. 132–134 (pp. 35, 38).

Capuano, A. et al. (2004). “Adverse drug events in two emergency departments in Naples, Italy: an observational study”. In: *Pharmacol. Res.* 50.6, pp. 631–636 (p. 24).
Carlson, Andrew et al. (2010). “Toward an Architecture for Never-Ending Language Learning.” In: *AAAI*. Vol. 5, p. 3 (p. 7).
Chambers, Nathanael and Dan Jurafsky (2011). “Template-based information extraction without the templates”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 976–986 (p. 44).
Chang, Chih-Chung and Chih-Jen Lin (2011). “LIBSVM: A library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27:1–27:27 (p. 19).
Chen, Jinxiu et al. (2006). “Relation extraction using label propagation based semi-supervised learning”. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 129–136 (p. 5).
Choi, Jinho D and Martha Palmer (2011). “Getting the most out of transition-based dependency parsing”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers – Vol. 2*. Association for Computational Linguistics, pp. 687–692 (p. 14).
Classen, David C et al. (2011). “âĂŸGlobal trigger toolâĂŹshows that adverse events in hospitals may be ten times greater than previously measured”. In: *Health Affairs* 30.4, pp. 581–589 (p. 39).
Clyne, Mindy et al. (2014). “Horizon scanning for translational genomic research beyond bench to bedside”. In: *Genetics in Medicine* 16.7, pp. 535–538 (p. viii).
Collis, J. (2010). “Adverse effects of overcrowding on patient experience and care”. In: *Emerg Nurse* 18.8, pp. 34–39 (p. 38).
Craven, M. and J. Kumlien (1999). “Constructing biological knowledge bases by extracting information from text sources”. In: *Proc Int Conf Intell Syst Mol Biol*, pp. 77–86 (p. 6).
Culotta, Aron and Jeffrey Sorensen (2004). “Dependency tree kernels for relation extraction”. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 423 (pp. 3, 4).
Deng, Lingjia and Janyce Wiebe (2014). “Sentiment propagation via implicature constraints”. In: *Proceedings of EACL* (p. 9).
Derry, Sheena, Yoon Kong Loke, and Jeffrey K Aronson (2001). “Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials”. In: *BMC Medical Research Methodology* 1.1, p. 7 (pp. vii, viii, 1).
Druss, B. G. and S. C. Marcus (2005). “Growth and decentralization of the medical literature: implications for evidence-based medicine”. In: *J Med Libr Assoc* 93.4, pp. 499–501 (p. vii).
Duke, Jon D and Jeff Friedlin (2010). “ADESSA: a real-time decision support service for delivery of semantically coded adverse drug event data”. In: *AMIA Annual Symposium Proceedings*. Vol. 2010. American Medical Informatics Association, p. 177 (pp. viii, 39).
Fan, Rong-En et al. (2008). “LIBLINEAR: A library for large linear classification”. In: *The Journal of Machine Learning Research* 9, pp. 1871–1874 (p. 44).
Filatova, Elena, Vasileios Hatzivassiloglou, and Kathleen McKeown (2006). “Automatic creation of domain templates”. In: *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, pp. 207–214 (p. 44).

Friedman, Carol (2000). “A broad-coverage natural language processing system.” In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 270 (pp. 35, 39).
— (2009). “Discovering novel adverse drug events using natural language processing and mining of the electronic health record”. In: *Artificial Intelligence in Medicine*. Springer, pp. 1–5 (p. 34).
Fundel, K., R. Kuffner, and R. Zimmer (2007). “RelEx–relation extraction using dependency parse trees”. In: *Bioinformatics* 23.3, pp. 365–371 (p. 3).
Furberg, C. D. and B. Pitt (2001). “Withdrawal of cerivastatin from the world market”. In: *Curr Control Trials Cardiovasc Med* 2.5, pp. 205–207 (p. vii).
Galanter, W. L., R. J. Didomenico, and A. Polikaitis (2005). “A Trial of Automated Decision Support Alerts for Contraindicated Medications Using Computerized Physician Order Entry.” In: *J Am Med Inform Assoc* 12.3, pp. 269–274 (p. 38).
Gandhi, T. K. et al. (2006). “Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims”. In: *Ann. Intern. Med.* 145.7, pp. 488–496 (p. 24).
Gandhi, T. K. et al. (2010). “Outpatient adverse drug events identified by screening electronic health records”. In: *J Patient Saf* 6.2, pp. 91–96 (p. 35).
Gardner, Matt and Tom Mitchell (2015). “Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction”. In: *Proceedings of EMNLP* 3 (p. 9).
Gish, Herbert (1990). “A probabilistic approach to the understanding and training of neural network classifiers”. In: *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, pp. 1361–1364 (p. 16).
Glasgow, Russell E and Karen M Emmons (2007). “How Can We Increase Translation of Research into Practice? Types of Evidence Needed”. In: *Annu. Rev. Public Health* 28, pp. 413–433 (p. viii).
GuoDong, Zhou, Qian LongHua, and Zhu QiaoMing (2009). “Label propagation via bootstrapped support vectors for semantic relation extraction between named entities”. In: *Computer Speech & Language* 23.4, pp. 464–478 (p. 5).
GuoDong, Zhou et al. (2005). “Exploring Various Knowledge in Relation Extraction”. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 427–434 (p. 19).
Gurulingappa, Harsha, Abdul Mateen-Rajput, Luca Toldo, et al. (2012). “Extraction of potential adverse drug events from medical case reports”. In: *Journal of Biomedical Semantics* 3.1, p. 15 (p. 35).
Gurwitz, J. H. et al. (2008). “Effect of computerized provider order entry with clinical decision support on adverse drug events in the long-term care setting”. In: *J Am Geriatr Soc* 56.12, pp. 2225–2233 (p. 38).
Hanbury, Allan (2012). “Medical Information Retrieval: An Instance of Domain-specific Search”. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’12. New York, NY, USA: ACM, pp. 1191–1192 (p. viii).
Hashimoto, Kazuma et al. (2013). “Simple Customization of Recursive Neural Networks for Semantic Relation Classification”. In: *EMNLP*, pp. 1372–1376 (p. 4).
Haug, P. J. et al. (2007). “Clinical decision support at Intermountain Healthcare”. In: *Clinical Decision Support Systems, Theory and Practice*. Ed. by E. S. Berner. Vol. 2. New York: Springer, pp. 159–189 (p. 35).
Hearst, Marti A (1992). “Automatic acquisition of hyponyms from large text corpora”. In: *Proceedings of the 14th Conference on Computational Linguistics-Volume 2*. Association for Computational Linguistics, pp. 539–545 (p. 43).

Herlocker, Jonathan L and Joseph A Konstan (2001). “Content-independent task-focused recommendation”. In: *Internet Computing, IEEE* 5.6, pp. 40–47 (p. 48).
Hermjakob, Henning et al. (2004). “IntAct: an open source molecular interaction database”. In: *Nucleic acids research* 32.suppl 1, pp. D452–D455 (p. 6).
Hill, R. G., L. M. Sears, and S. W. Melanson (2013). “4000 clicks: a productivity analysis of electronic medical records in a community hospital ED”. In: *Am J Emerg Med* 31.11, pp. 1591–1594 (p. 38).
Hodges, P. E., W. E. Payne, and J. I. Garrels (1998). “The Yeast Protein Database (YPD): a curated proteome database for Saccharomyces cerevisiae”. In: *Nucleic Acids Res.* 26.1, pp. 68–72 (p. 6).
Hoffmann, Raphael et al. (2011). “Knowledge-based weak supervision for information extraction of overlapping relations”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 541–550 (pp. 3, 5, 6, 19).
Hohl, C. M. et al. (2010). “Do emergency physicians attribute drug-related emergency department visits to medication-related problems?” In: *Ann Emerg Med* 55.6, pp. 493–502 (p. 37).
Honigman, B. et al. (2001). “Using computerized data to identify adverse drug events in outpatients”. In: *J Am Med Inform Assoc* 8.3, pp. 254–266 (p. 35).
Horn, John R, Philip D Hansten, and Lingtak-Neander Chan (2007). “Proposal for a new tool to evaluate drug interaction cases”. In: *Annals of Pharmacotherapy* 41.4, pp. 674–680 (p. 38).
Hristovski, D., T. Rindflesch, and B. Peterlin (2013). “Using literature-based discovery to identify novel therapeutic approaches”. In: *Cardiovasc Hematol Agents Med Chem* 11.1, pp. 14–24 (p. 10).
Hristovski, D. et al. (2006). “Exploiting semantic relations for literature-based discovery”. In: *AMIA Annu Symp Proc*, pp. 349–353 (p. 10).
Izquierdo-Garcia, E. and I. Escobar-Rodriguez (2012). “[Systematic review of new protease inhibitors interactions: telaprevir and boceprevir]”. In: *Farm Hosp* 36.6, pp. 469–482 (p. 35).
Jaccard, Paul (1901). *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge (p. 31).
Jayarama, N, K Shiju, and K Prabahakar (2012). “Adverse drug reactions in adults leading to emergency department visits”. In: *Int J Pharm Pharm Sci* 4, pp. 642–646 (pp. ix, 37).
Jensen, David and Jennifer Neville (2002). “Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning”. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. ICML ’02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 259–266 (p. 7).
Jha, A. K. et al. (2009). “Use of electronic health records in U.S. hospitals”. In: *N. Engl. J. Med.* 360.16, pp. 1628–1638 (p. 38).
Jiang, G., H. R. Solbrig, and C. G. Chute (2011). “ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology”. In: *AMIA Annu Symp Proc* 2011, pp. 607–616 (p. 39).
Kemp, Charles et al. (2006). “Learning Systems of Concepts with an Infinite Relational Model”. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*. AAAI’06. Boston, Massachusetts: AAAI Press, pp. 381–388 (p. 8).
Kim, Jin-Dong et al. (2011). “Overview of BioNLP shared task 2011”. In: *Proceedings of the BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics, pp. 1–6 (p. 3).
Knox, C. et al. (2011). “DrugBank 3.0: a comprehensive resource for ’omics’ research on drugs”. In: *Nucleic Acids Res.* 39.Database issue, pp. D1035–1041 (pp. viii, 12, 43).

Kok, Stanley and Pedro Domingos (2007). “Statistical predicate invention”. In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 433–440 (p. 8).

Komachi, Mamoru et al. (2008). “Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1011–1020 (p. 15).

Koppel, R. et al. (2008). “Identifying and quantifying medication errors: evaluation of rapidly discontinued medication orders submitted to a computerized physician order entry system”. In: *J Am Med Inform Assoc* 15.4, pp. 461–465 (pp. 25, 31).

Kullback, Solomon and Richard A Leibler (1951). “On information and sufficiency”. In: *The annals of mathematical statistics*, pp. 79–86 (p. 17).

Laan, Anna Laura van der and Marianne Boenink (2012). “Beyond bench and bedside: disentangling the concept of translational research”. In: *Health care analysis* 23.1, pp. 32–49 (p. viii).

Lao, Ni, Tom Mitchell, and William W. Cohen (2011). “Random Walk Inference and Learning in a Large Scale Knowledge Base”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, pp. 529–539 (p. 8).

Lee, E. K. et al. (2010). “Improving Patient Safety through Medical Alert Management: An Automated Decision Tool to Reduce Alert Fatigue”. In: *AMIA Annu Symp Proc* 2010, pp. 417–421 (p. 39).

Lenfant, Claude (2003). “Clinical Research to Clinical Practice – Lost in Translation?” In: *New England Journal of Medicine* 349.9, pp. 868–874 (p. viii).

Leroy, G. et al. (2008). “Evaluating online health information: beyond readability formulas”. In: *AMIA Annu Symp Proc*, pp. 394–398 (p. 1).

Li, Jiao and Zhiyong Lu (2013). “Pathway-based drug repositioning using causal inference”. In: *BMC bioinformatics* 14.Suppl 16, S3 (p. viii).

Lialiou, Pascalina and John Mantas (2014). “Online Information Retrieval Systems and Health Professionals”. In: *Integrating Information Technology and Management for Quality of Care* 202, pp. 146–148 (p. viii).

Lindberg, D., B. Humphreys, and A. McCray (1993). “The Unified Medical Language System”. In: *Methods of Information in Medicine* 32.4, pp. 281–291 (pp. viii, 2, 6, 26, 40).

Lipczak, Marek, Arash Koushkestani, and Evangelos Milios (2014). “Tulip: lightweight entity recognition and disambiguation using wikipedia-based topic centroids”. In: *Proceedings of the first international workshop on Entity recognition & disambiguation*. ACM, pp. 31–36 (p. 48).

Liu, Xiao and Hsinchun Chen (2013). “AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums”. In: *Smart Health*. Springer, pp. 134–150 (p. 3).

Liu, Yudong, Zhongmin Shi, and Anoop Sarkar (2007). “Exploiting rich syntactic information for relation extraction from biomedical articles”. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Association for Computational Linguistics, pp. 97–100 (p. 19).

Lodhi, Huma et al. (2002). “Text classification using string kernels”. In: *The Journal of Machine Learning Research* 2, pp. 419–444 (p. 4).

Luciano, Joanne S et al. (2011). “The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside.” In: *J. Biomedical Semantics* 2.S-2, S1 (p. 1).

Mannheimer, Buster (2009). “Drug-related problems with special emphasis on drug-drug interactions”. In: Institutionen för klinisk forskning och utbildning, Södersjukhuset/Department of Clinical Science and Education, Södersjukhuset, pp. 7–8 (p. 38).

Marrero, Mónica et al. (2012). “Information retrieval systems adapted to the biomedical domain”. In: *arXiv preprint arXiv:1203.6845* (p. 2).

Martins, André FT, Miguel Almeida, and Noah A Smith (2013). “Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers.” In: *ACL (2)*. Association for Computational Linguistics, pp. 617–622 (p. 2).

Maslennikov, Mstislav and Tat-Seng Chua (2007). “A multi-resolution framework for information extraction from free text”. In: *Annual Meeting of the Association for Computational Linguistics*. Vol. 45. 1, p. 592 (p. 43).

McCarthy, Diana et al. (2004). “Finding predominant word senses in untagged text”. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 279 (p. 17).

McIntosh, Tara and James R Curran (2009). “Reducing semantic drift with bagging and distributional similarity”. 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: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 396–404 (pp. 5, 15).

Melton, G. B. and G. Hripcsak (2005). “Automated detection of adverse events using natural language processing of discharge summaries”. In: *J Am Med Inform Assoc* 12.4, pp. 448–457 (pp. 35, 38).

Mihalcea, Rada, Courtney Corley, and Carlo Strapparava (2006). “Corpus-based and knowledge-based measures of text semantic similarity”. In: *AAAI*. Vol. 6, pp. 775–780 (p. 17).

Miller, George A. (1995). “WordNet: A Lexical Database for English”. In: *Commun. ACM* 38.11, pp. 39–41 (pp. 17, 30, 47).

Mintz, Mike et al. (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: Volume 2-Volume 2*. Association for Computational Linguistics, pp. 1003–1011 (pp. 5, 6, 19).

Mooney, Raymond J and Razvan C Bunescu (2005). “Subsequence kernels for relation extraction”. In: *Advances in neural information processing systems*, pp. 171–178 (pp. 3, 4).

Murff, H. J. et al. (2003). “Detecting adverse events for patient safety research: a review of current methodologies”. In: *J Biomed Inform* 36.1-2, pp. 131–143 (p. 35).

Musto, Cataldo et al. (2014). “Combining distributional semantics and entity linking for context-aware content-based recommendation”. In: *User Modeling, Adaptation, and Personalization*. Springer, pp. 381–392 (p. 48).

Nédellec, Claire et al. (2013). “Overview of BioNLP shared task 2013”. In: *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 1–7 (p. 3).

Neelakantan, Arvind, Benjamin Roth, and Andrew McCallum (2015). “Compositional Vector Space Models for Knowledge Base Completion”. In: *Association for Computational Linguistics (ACL)* (p. 9).

Nguyen, Truc-Vien T, Alessandro Moschitti, and Giuseppe Riccardi (2009). “Convolution kernels on constituent, dependency and sequential structures for relation extraction”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, pp. 1378–1387 (p. 3).

Nickel, Maximilian (2013). “Tensor Factorization for Relational Learning”. PhD thesis. Ludwig Maximilian University of Munich (pp. 7, 8).

Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel (2011). "A three-way model for collective learning on multi-relational data". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 809–816 (p. 8).

Nickel, Maximilian et al. (2015). "A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction". In: *arXiv preprint arXiv:1503.00759* (p. 7).

Niculescu-Mizil, Alexandru and Rich Caruana (2005). "Predicting good probabilities with supervised learning". In: *Proceedings of the 22nd international conference on Machine learning*. ACM, pp. 625–632 (p. 8).

Nikfarjam, A. et al. (2015). "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features". In: *J Am Med Inform Assoc* 22.3, pp. 671–681 (p. 3).

Null, Gary et al. (2005). "Death by medicine". In: *Journal of Orthomolecular Medicine* 20.1, pp. 21–34 (p. 37).

Olshaker, J. S. and N. K. Rathlev (2006). "Emergency Department overcrowding and ambulance diversion: the impact and potential solutions of extended boarding of admitted patients in the Emergency Department". In: *J Emerg Med* 30.3, pp. 351–356 (p. 38).

Patwardhan, Siddharth and Ellen Riloff (2007). "Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions." In: *EMNLP-CoNLL*. Vol. 7, pp. 717–727 (p. 43).

— (2009). "A unified model of phrasal and sentential evidence for information extraction". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 151–160 (p. 44).

Percha, B., Y. Garten, and R. B. Altman (2012). "Discovery and explanation of drug-drug interactions via text mining". In: *Pac Symp Biocomput*, pp. 410–421 (p. 35).

Phansalkar, S. et al. (2013). "Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records". In: *J Am Med Inform Assoc* 20.3, pp. 489–493 (p. 39).

Platt, John C. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3, pp. 61–74 (pp. 8, 14).

Pratt, Wanda and Meliha Yetisgen-Yildiz (2003). "LitLinker: capturing connections across the biomedical literature". In: *Proceedings of the 2nd international conference on Knowledge capture*. ACM, pp. 105–112 (p. 10).

Ramesh, B. P. et al. (2014). "Automatically Recognizing Medication and Adverse Event Information From Food and Drug Administration's Adverse Event Reporting System Narratives". In: *JMIR Med Inform* 2.1, e10 (p. 39).

Ramos, K., R. Linscheid, and S. Schafer (2003). "Real-time information-seeking behavior of residency physicians". In: *Fam Med* 35.4, pp. 257–260 (p. 38).

Ratain, M. J. and W. K. Plunkett (2003). "Principles of Pharmacodynamics". In: *Holland-Frei Cancer Medicine*. Ed. by D. W. Kufe et al. 6th ed. BC Decker (p. 11).

Reaume, Andrew G (2012). "Drug repurposing through nonhypothesis driven phenotypic screening". In: *Drug Discovery Today: Therapeutic Strategies* 8.3, pp. 85–88 (p. viii).

Richard, Michael D and Richard P Lippmann (1991). "Neural network classifiers estimate Bayesian a posteriori probabilities". In: *Neural computation* 3.4, pp. 461–483 (p. 16).

Riedel, Sebastian, Limin Yao, and Andrew McCallum (2010). "Modeling relations and their mentions without labeled text". In: *Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 148–163 (p. 6).

Rindflesch, T. C. and M. Fiszman (2003). “The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text”. In: *J Biomed Inform* 36.6, pp. 462–477 (p. 10).

Rodriguez-Terol, A. et al. (2009). “Quality of interaction database management systems”. In: *Farm Hosp* 33.3, pp. 134–146 (pp. vii, 1).

Rogers, F. B. (1963). “Medical subject headings”. In: *Bull Med Libr Assoc* 51, pp. 114–116 (pp. 30, 40).

Roller, Roland and Mark Stevenson (2014). “Applying UMLS for Distantly Supervised Relation Detection”. In: *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pp. 80–84 (p. 6).

Roulet, L. et al. (2014). “Adverse drug event nonrecognition in emergency departments: an exploratory study on factors related to patients and drugs”. In: *J Emerg Med* 46.6, pp. 857–864 (p. 37).

Rozich, JD, CR Haraden, and RK Resar (2003). “Adverse drug event trigger tool: a practical methodology for measuring medication related harm”. In: *Quality and Safety in Health Care* 12.3, pp. 194–200 (p. 39).

Santos, Cıcero Nogueira dos, Bing Xiang, and Bowen Zhou (2015). “Classifying relations by ranking with convolutional neural networks”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 626–634 (p. 4).

Sarker, A. and G. Gonzalez (2015). “Portable automatic text classification for adverse drug reaction detection via multi-corpus training”. In: *J Biomed Inform* 53, pp. 196–207 (p. 39).

Schedlbauer, A. et al. (2009). “What evidence supports the use of computerized alerts and prompts to improve clinicians’ prescribing behavior?” In: *J Am Med Inform Assoc* 16.4, pp. 531–538 (p. 38).

Schwartz, A. S. and M. A. Hearst (2003). “A simple algorithm for identifying abbreviation definitions in biomedical text”. In: *Proceedings of the 8th Pacific Symposium on Biocomputing*, pp. 451–462 (pp. 13, 42, 47).

Scorza, K. et al. (2007). “Evaluation of nausea and vomiting”. In: *Am Fam Physician* 76.1, pp. 76–84 (p. 31).

Seger, Andrew C et al. (2005). *Development of a computerized adverse drug event (ADE) monitor in the outpatient setting*. Tech. rep. DTIC Document (p. 35).

Segura-Bedmar, I., P. Martinez, and C. de Pablo-Sanchez (2011). “Using a shallow linguistic kernel for drug-drug interaction extraction”. In: *J Biomed Inform* 44.5, pp. 789–804 (p. 35).

Segura-Bedmar, I., P. Martínez, and D. Sánchez-Cisneros (2011). “The 1st DDI Extraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts”. In: *Proc. DDI Extraction-2011 challenge task*, 1–âĂŞ9 (p. 35).

Segura Bedmar, Isabel, Paloma Martínez, and María Herrero Zazo (2013). “Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)”. In: Association for Computational Linguistics (p. 3).

Seydel, JK and K-J Schaper (1981). “Quantitative structure-pharmacokinetic relationships and drug design”. In: *Pharmacology & therapeutics* 15.2, pp. 131–182 (p. 11).

Shetty, Kanaka and Siddhartha R. Dalal (2011). “Using information mining of the medical literature to improve drug safety”. In: *J. Am. Med. Inform. Assoc.* 18, pp. 668–674 (p. 35).

Sijs, H. van der et al. (2006). “Overriding of drug safety alerts in computerized physician order entry”. In: *J Am Med Inform Assoc* 13.2, pp. 138–147 (pp. 31, 39).

Sijs, H. van der et al. (2008). “Turning off frequently overridden drug alerts: limited opportunities for doing it safely”. In: *J Am Med Inform Assoc* 15.4, pp. 439–448 (p. 39).

Singh, H. (2013). “Diagnostic errors: moving beyond ’no respect’ and getting ready for prime time”. In: *BMJ Qual Saf* 22.10, pp. 789–792 (pp. ix, 24, 32).

Smalheiser, N. R. and D. R. Swanson (1998). “Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses”. In: *Comput Methods Programs Biomed* 57.3, pp. 149–153 (p. 10).

Smith, L., T. Rindflesch, and W. J. Wilbur (2004). “MedPost: A Part of Speech Tagger for BioMedical Text”. In: *Bioinformatics* 20.14, pp. 2320–2321 (p. 27).

Smithburger, P. L., S. L. Kane-Gill, and A. L. Seybert (2012). “Drug-drug interactions in the medical intensive care unit: an assessment of frequency, severity and the medications involved”. In: *Int J Pharm Pract* 20.6, pp. 402–408 (pp. 35, 36).

Socher, Richard et al. (2012). “Semantic compositionality through recursive matrix-vector spaces”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 1201–1211 (p. 4).

Strom, B. L. et al. (2010). “Unintended effects of a computerized physician order entry nearly hard-stop alert to prevent a drug interaction: a randomized controlled trial”. In: *Arch. Intern. Med.* 170.17, pp. 1578–1583 (p. 38).

Suchanek, Fabian M, Gjergji Kasneci, and Gerhard Weikum (2007). “Yago: a core of semantic knowledge”. In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706 (p. 7).

Sun, Ang, Ralph Grishman, and Satoshi Sekine (2011). “Semi-supervised relation extraction with large-scale word clustering”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 521–529 (p. 5).

Surdeanu, Mihai et al. (2012). “Multi-instance multi-label learning for relation extraction”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 455–465 (p. 6).

Swanson, D. R. (1986). “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge”. In: *Perspect. Biol. Med.* 30.1, pp. 7–18 (pp. 10, 15).

Tatonetti, Nicholas P., Guy Haskin Fernald, and Russ B. Altman (2011). “A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports”. In: *J. Am. Med. Inform. Assoc.* 19.1, pp. 79–85 (pp. 35, 39).

Tatonetti, Nicholas P et al. (2012). “Data-driven prediction of drug effects and interactions”. In: *Science translational medicine* 4.125, 125ra31–125ra31 (pp. viii, 39).

Théophile, Hélène et al. (2012). “An updated method improved the assessment of adverse drug reaction in routine pharmacovigilance”. In: *Journal of clinical epidemiology* 65.10, pp. 1069–1077 (p. 35).

Thomas, Philippe et al. (2011). “Learning protein protein interaction extraction using distant supervision”. In: *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pp. 34–41 (p. 6).

Trifiro, G. et al. (2005). “Adverse drug events in emergency department population: a prospective Italian study”. In: *Pharmacoepidemiol Drug Saf* 14.5, pp. 333–340 (pp. 24, 37).

Trzeciak, S. and E. P. Rivers (2003). “Emergency department overcrowding in the United States: an emerging threat to patient safety and public health”. In: *Emerg Med J* 20.5, pp. 402–405 (p. 38).

Tsai, H. H. et al. (2013). "A review of potential harmful interactions between anticoagulant/antiplatelet agents and Chinese herbal medicines". In: *PLoS ONE* 8.5, e64255 (p. 35).
Tsuruoka, Yoshimasa et al. (2005). "Developing a robust part-of-speech tagger for biomedical text". In: *Advances in informatics*. Springer, pp. 382–392 (pp. 13, 28, 46).
Turian, Joseph, Lev Ratinov, and Yoshua Bengio (2010). "Word representations: a simple and general method for semi-supervised learning". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, pp. 384–394 (p. 4).
UMLS® Reference Manual [Internet] (2009). National Library of Medicine (US). Bethesda (MD) (pp. 13, 28).
Voorhees, Ellen M et al. (1999). "The TREC-8 Question Answering Track Report". In: *TREC*. Vol. 99, pp. 77–82 (pp. 33, 52).
Wang, Chang and James Fan (2014). "Medical relation extraction with manifold models". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 828–838 (pp. 10, 12).
Wang, W. et al. (2011). "A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations". In: *AMIA Annu Symp Proc* 2011, pp. 1464–1470 (p. 35).
Wang, X. et al. (2009). "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study". In: *J Am Med Inform Assoc* 16.3, pp. 328–337 (pp. 35, 39).
Weeber, Marc et al. (2000). "Text-based discovery in biomedicine: the architecture of the DAD-system." In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, p. 903 (p. 2).
Weissman, J. S. et al. (2007). "Hospital workload and adverse events". In: *Med Care* 45.5, pp. 448–455 (p. 38).
Woolf, Steven H (2008). "The meaning of translational research and why it matters". In: *Journal of the American Medical Association* 299.2, pp. 211–213 (p. 1).
World Health Organization, (2006). "The safety of medicines in public health programmes: pharmacovigilance, an essential tool". In: (p. 34).
Wu, D. T. et al. (2015). "Assessing the readability of clinicaltrials.gov". In: *J Am Med Inform Assoc* (pp. 1, 2).
Wu, Zhibiao and Martha Palmer (1994). "Verbs semantics and lexical selection". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 133–138 (pp. 17, 31).
Xu, Feiyu, Hans Uszkoreit, and Hong Li (2007). "A seed-driven bottom-up machine learning framework for extracting relations of various complexity". In: *ACL*. Vol. 7, pp. 584–591 (p. 3).
Xu, R. and Q. Wang (2013). "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing". In: *BMC Bioinformatics* 14, p. 181 (p. 3).
Xu, Wei et al. (2013). "Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction." In: *ACL (2)*, pp. 665–670 (p. 3).
Xu, Zhao et al. (2012). "Infinite hidden relational models". In: *arXiv preprint arXiv:1206.6864* (p. 8).
Yakushiji, Akane et al. (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*. Association for Computational Linguistics, pp. 284–292 (p. 3).

Yang, Bishan et al. (2015). “Embedding Entities and Relations for Learning and Inference in Knowledge Bases”. In: *International Conference on Learning Representations* (p. 8).

Yao, Limin, Sebastian Riedel, and Andrew McCallum (2010). “Collective Cross-document Relation Extraction Without Labelled Data”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’10. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 1013–1023 (p. 8).

Yarowsky, David (1995). “Unsupervised word sense disambiguation rivaling supervised methods”. In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 189–196 (p. 5).

Yeleswarapu, S. et al. (2014). “A pipeline to extract drug-adverse event pairs from multiple data sources”. In: *BMC Med Inform Decis Mak* 14, p. 13 (p. 39).

Zed, P. J. et al. (2008). “Incidence, severity and preventability of medication-related visits to the emergency department: a prospective study”. In: *CMAJ* 178.12, pp. 1563–1569 (pp. ix, 24, 37).

Zelenko, Dmitry, Chinatsu Aone, and Anthony Richardella (2003). “Kernel methods for relation extraction”. In: *The Journal of Machine Learning Research* 3, pp. 1083–1106 (pp. 3, 4).

Zeng, Daojian et al. (2014). “Relation classification via convolutional deep neural network”. In: *Proceedings of COLING*, pp. 2335–2344 (p. 4).

Zeng, Daojian et al. (2015). “Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks”. In: EMNLP (p. 4).

Zeng-Treitler, Q. et al. (2007). “Text characteristics of clinical reports and their implications for the readability of personal health records”. In: *Stud Health Technol Inform* 129.Pt 2, pp. 1117–1121 (p. 1).

Zerhouni, Elias A (2005). “Translational and clinical science – time for a new vision”. In: *New England Journal of Medicine* 353.15, pp. 1621–1623 (p. 1).

Zhang, Zhu (2004). “Weakly-supervised relation classification for information extraction”. In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, pp. 581–588 (pp. 5, 19).

Zhi, M. et al. (2013). “The landscape of inappropriate laboratory testing: a 15-year meta-analysis”. In: *PLoS ONE* 8.11, e78962 (pp. ix, 24, 32).