Bridging semantics and syntax with graph algorithms—state-of-the-art of extracting biomedical relations

Yuan Luo, Özlem Uzuner and Peter Szolovits

Abstract

Research on extracting biomedical relations has received growing attention recently, with numerous biological and clinical applications including those in pharmacogenomics, clinical trial screening and adverse drug reaction detection. The ability to accurately capture both semantic and syntactic structures in text expressing these relations becomes increasingly critical to enable deep understanding of scientific papers and clinical narratives. Shared task challenges have been organized by both bioinformatics and clinical informatics communities to assess and advance the state-of-the-art research. Significant progress has been made in algorithm development and resource construction. In particular, graph-based approaches bridge semantics and syntax, often achieving the best performance in shared tasks. However, a number of problems at the forefronts of biomedical relation extraction continue to pose interesting challenges and present opportunities for great improvement and fruitful research. In this article, we place biomedical relation extraction against the backdrop of its versatile applications, present a gentle introduction to its general pipeline and shared resources, review the current state-of-the-art in methodology advancement, discuss limitations and point out several promising future directions.

Key words: biomedical relation extraction; natural language processing; graph mining; machine learning; scientific literature; clinical narratives

Introduction

Relation extraction from text documents is an important task in knowledge representation and inference to create or augment structured knowledge bases and in turn support question answering and decision making. The task generally involves annotating unstructured text with named entities and identifying the relations between these annotated entities. State-of-the-art named entity recognizers can automatically annotate text with high accuracy [1, 2], but relation extraction is not as straightforward. General domain relation extraction is an active research area for decades [3]. In the biomedical and clinical domain, extracting relations from scientific publications and clinical narratives has been gaining traction over the past decade, and is the focus of this review.

To illustrate the importance of biomedical relation extraction, consider that in lymphoma pathology reports, immunophenotypic features are expressed as relations among medical concepts. For example, in ‘[large atypical cells] are positive for [CD30] and negative for [CD15]’, ‘large atypical cells’, ‘CD30’ and ‘CD15’ are medical concepts; ‘CD30’ and ‘CD15’ are cell surface antigens. A bag-of-words or bag-of-concepts representation would fail to capture whether ‘large atypical cells’ are positive...
The representation of relations has been a subject of knowledge representation research for decades [5], with various alternatives. One representation uses composed simple logical forms. For example, Resource Description Framework (RDF) or Web Ontology Language encodes complex relations by multiple triples, where the elements of these triples can themselves be other composed forms. Thus, binary relations such as positive-expression (large atypical cells, CD30) have the following subject-predicate-object triple representation: large atypical cells-positively express-CD30. A more powerful alternative is the sentential logic (or propositional logic) representation [5], in which relations are propositions or composed propositions using logical connectives (e.g. ‘and’ for conjunction, ‘or’ for disjunction). Compared with RDF triples, propositional logic has additional constructs such as connectives and inference rules, thus is more expressive. A third alternative is the graph-based representation in which nodes are named entities and edges indicate relationships, or multiple named entities connected by multiple edges can be regarded as one relation, as shown in Figure 1.

This review focuses on NLP methods using graph-based representations and algorithms to extract biomedical relations from unstructured text. Regarding alternative representations, the sentential logic representation can be noted with graph-based representation [6]. Biomedical relations (including events) can be universally represented as graphs by converting biomedical concepts to nodes and syntactic/semantic links to edges. Other propositional representations may require specific interpretation of the graphs. For instance, representing the negation of a proposition may require the introduction of nested graphs, and to give special semantics to a relation labeled NOT. Furthermore, although composition leads to complexity (e.g. n-ary relations or nested relations), by adopting a graph-based representation, we can focus on common syntactic and semantic graphical patterns that provide good ways to capture relations. In fact, as will become clear later in this review, almost all state-of-the-art methods for extracting relations are graph-based. This article assumes that the reader has a basic knowledge of biomedical NLP; see [7-12] for introductions and surveys on latest applications.

The reader should also be aware of a body of research on curated structured knowledge bases, with manual annotations of biomedical relations by experts. Some of these knowledge bases are biologically focused, such as KEGG [13], STRING [14], InterPro [15] and InterDom [16]. Others are more clinically focused, such as PharmGKB [17], VARMED [18] and ClinVar [19]. However, the expert sourcing methods often scale poorly with the exponentially growing body of biomedical free text. Thus, automated methods present a promising direction for discovering relations to augment existing knowledge bases; this motivated many methods discussed in this article.

**Application of biomedical relation extraction**

Extraction biomedical relations has numerous applications that vary from advancing basic sciences to improving clinical practices, as shown in Figure 2. These applications include but are not limited to biomolecular information extraction, clinical trial screening, pharmacogenomics, diagnosis categorization, as well as discovery of adverse drug reactions (ADRs) and drug–drug interactions (DDIs).
known disease–gene relations from the literature, which are genes for further validation from a large number of candidates mining has shown its value in the prioritization of cancerous collectively call biomolecular information extraction. Such relation gene ontology [24] and pathway information [25], which we col-
lected methods have been applied to mining protein–protein interactions (PPIs) [20, 21], gene–phenotype associations [22, 23], gene ontology [24] and pathway information [25], which we collectively call biomolecular information extraction. Such relation mining has shown its value in the prioritization of cancerous genes for further validation from a large number of candidates [26]. Many of these approaches apply NLP methods to extract known disease–gene relations from the literature, which are then used to predict novel disease–gene relations [27–31].

Biomolecular information extraction

To keep up with the exponential growth of the literature, automated methods have been applied to mining protein–protein interactions (PPIs) [20, 21], gene–phenotype associations [22, 23], gene ontology [24] and pathway information [25], which we collectively call biomolecular information extraction. Such relation mining has shown its value in the prioritization of cancerous genes for further validation from a large number of candidates [26]. Many of these approaches apply NLP methods to extract known disease–gene relations from the literature, which are then used to predict novel disease–gene relations [27–31].

Clinical trial screening

 Archived clinical and research data have been made available by governmental agencies and corporations, such as ClinicalTrials.gov [32]. Clinical trials are in large part characterized by eligibility criteria, some of which can be captured via relations (e.g. no [diagnosis] for [rheumatoid arthritis] for at least [6 months]). Electronic screening can improve efficiency in clinical trial recruitment, and intelligent query over trials can support clinical knowledge curation [33]. Recently, NLP support has proved useful in automatically detecting named entities in eligibility criteria [34, 35], and further in extracting relations between named entities to characterize eligibility criteria [36–38].

Pharmacogenomics

Pharmacogenomics aims to understand how different patients respond to drugs by studying relations between drug response phenotypes and patient genetic variations. Much of this knowledge can be mined from scientific literature and curated in databases to enable discovering new relationships. One such database is the Pharmacogenetics Research Network and Knowledge Base (PharmGKB [39]). Initial efforts to populate PharmGKB included a mixture of expert annotation and rule-based approaches. Recent approaches have extended to using semantic and syntactic analysis as well as statistical machine learning tools to mine targeted pharmacogenomics relations from scientific literature and clinical records [40–42].

Diagnosis categorization

Diagnosis categorization enables automated billing and patient cohort selection for secondary research. Developed systems can automatically code and classify diagnoses from Electronic Medical Records (EMRs) [43–47]. Recent work demonstrated the success of extracting semantic relations and using these relations as additional features in diagnosis categorization, some through better grouping features using curated relations [48], others through unsupervised learning to extract more expressive representation of relations between medical concepts [49, 50].

ADR and DDI

ADR refers to unexpected injuries caused by taking a medication. DDI happens when a drug affects the activity of another drug simultaneously administered. ADR is an important cause of morbidity and mortality [51], and DDIs may cause reduced drug efficacy or drug overdose. Detecting potential ADRs and DDIs can guide the process of drug development. An increasing number of methods have leveraged the scientific literature and clinical records using NLP. These systems often explore the relations between drugs, genes and pathways, and discover ADRs [52–54] and DDIs [55, 56] stated in text. A large amount of research in recent years also explored user-generated content in social networks to detect ADRs, see [57] for a recent review.

General pipeline for biomedical relation extraction

In Figure 3, we first present a general pipeline, summarized from the reviewed approaches, as a cookbook to follow either in part or as a whole for extracting biomedical relations. The reader can refer the components discussed in the state-of-the-art methods to this cookbook to gain comprehensive understanding. For completeness, we assume documents as input and extracted relations as output. The pipeline starts with section recognition, which distinguishes text under different section headings (e.g. ‘Chief Complaints’ or ‘Past Medical History’). Sentence breaking then automatically decides where sentences in a paragraph begin and end. Typographical analysis investigates features such as capitalization and usage of alphanumeric characters. Stemming reduces the inflected words to the root form (e.g. performed to perform). POS tagging assigns a part-of-speech tag for each word in the sentence (e.g. VBN for ‘per-
formed’ in the sentence in Figure 1). Parsing is the process of assigning a syntactic structure to a sentence (e.g. the constituency or dependency structure obtained by Stanford Parser). The results from typographical analysis, stemming, POS tagging and parsing can provide features for recognizing anaphora (coreference resolution) and typed concepts (concept recognition). Coreference resolution and concept resolution can also improve parsing accuracy. Together with parsing, they are essential in generating the graph representation for a sentence and labeling semantic roles of concepts in the graph representation (Semantic Role Labeling). The graph representation is the foundation for graph mining, and along with upstream steps including direct regular expression feature extraction, leads to the generation of semantically and syntactically enriched features.
These features then support either rule-based, feature-space-based or kernel-based relation extraction system. Many biomedical relation extraction systems rely on external knowledge sources [e.g. Unified Medical Language System (UMLS) [58]].

The pipeline can also be used as a foundation for downstream applications such as logical inference with extracted relations. The pipeline covers steps for breaking the documents to sentences, understanding the semantic and syntactic structures of sentences and constructing a multitude of features for relation extraction. We refer the reader to the description of each step in the caption of Figure 3. We emphasize the role of graph mining in the pipeline as a central concept. The mined graphs provide a converging point for methods that combine local features (e.g. tokens and part-of-speech tags), a diverging point where more integrated features (e.g. relations as features) are constructed, and a bridge to connect the syntax and semantics.

**Introduction to graph representations and graph algorithms for biomedical relation extraction**

Graph representation for narrative sentences naturally breaks down to the choice of information represented by nodes and edges. Common node choices include tokens (e.g. biopsy), named entities (e.g. bone marrow biopsy), semantically labeled named entity (e.g. bone marrow biopsy-Diagnostic Procedure, where Diagnostic Procedure is a UMLS semantic type) and relations themselves within nested relations. Those choices are in the order of encoding increasingly enriched and complex semantic information. Common edge choices include syntactic dependency (e.g. dobj for direct object, in dependency parsing), syntactic constituency link (e.g. NP-NN, in constituency parsing), event argument (e.g. event theme) and association (e.g. based on co-occurrence or customized statistics) as edge. These choices integrate semantic and syntactic information into the graph representation to different degrees. Graph mining algorithms can be applied to extract relations directly or construct useful features. Frequently used algorithms can be categorized as follows: some identify shortest path (or its variant) between concept pair, which can be performed using standard algorithms such as Dijkstra’s algorithm [59]; some create association graphs then try to apply customized labels to them; some use subgraph matching to compare the similarity between subgraphs based on a score that aggregates node distances and edge distances; some carry out frequent subgraph or subtree pattern mining to directly extract candidate relations; some...
State-of-the-art methods for biomedical relation extraction

Methodology for biomedical relation extraction has received increasing attention. Conventional approaches focus on using co-occurrence statistics as a proxy for relatedness [41, 102–105]. Some clinical NLP systems apply hand-crafted syntactic and semantic rules to extract prespecified semantic relations, such as MedLEE [106] and SemRep [107], but are hard to adapt to new subdomains. Recent research focuses more on syntactic parsing, to develop generalizable methods to extract relations that fully explore the constituency and dependency structures of natural language. In this section, we review the state-of-the-art approaches where graph (including tree) mining techniques are used to derive relations from syntactic or semantic parses. We categorize the methods according to whether their corpora mainly concern scientific publications or clinical narratives, as this content difference often has implications for the methods and resources used to extract relations. On the other hand, community challenges continuously help to promote the development of state-of-the-art methods. Thus, in each category, we organize research work around the challenge they participate or as non-challenge-participating research. We summarize the algorithms and systems in Table 1.

Relation extraction from the scientific literature

Over the past decade, continuous effort has been directed to extracting semantic relations from biomedical literature text, often in the form of shared-task community challenges that aim to assess and advance NLP techniques. Notable community challenges include BioNLP shared tasks on event mining, BioCreative shared tasks on PPI extraction and DDIExtraction challenges on DDI extraction. We observed that an increasing number of teams applied graph-based techniques to characterize the semantic relations in these shared tasks. These techniques are frequently placed among the top-performing echelon. This section reviews the graph-based methodologies developed for these challenges. We consider only the papers accepted into the shared task proceedings as full publications, and focus on the top-performing systems. We summarize the f-measures of the best systems in each shared task as a representative evaluation, and refer the reader to the challenge overviews for detailed and comprehensive evaluations. Perhaps through learning the lessons from these challenges, real-world applications such as the field of pharmacogenomics also saw significant momentum in development of graph-based text-mining methods. We devote the last part of this section to recent advances in pharmacogenomics and demonstrate the transfer and adaptation of graph-based algorithms from methodology-oriented research to application-oriented research in biomedical relation extraction.

BioNLP event extraction shared tasks

Three BioNLP shared tasks (ST) have focused on recognizing biological events (relations) from the literature. These tasks provided the protein mentions as input and asked the participating teams to identify a predefined set of semantic relations. Teams were not required to discover the protein mentions. BioNLP-ST-2009 consisted of three tasks, including core event detection, event argument recognition and negation/speculation detection, all based on the GENIA corpus [108]. BioNLP-ST-2011 expanded the tasks and resources to cover more text types, event types and subject domains [109]. Besides the continued GENIA task (GE), BioNLP-ST-2011 added the following tasks: epigenetics and post-translational modification (EPI), infectious diseases (ID), bacteria biotope (BB) and bacteria interaction (BI). BioNLP-ST-2013 further expanded the application domains with tasks of GE, BB, cancer genetics (CG), pathway curation (PC) and gene regulation ontology (GRO) [110]. Table 2 describes those tasks in more detail.

The typical event extraction workflow can be broken into two general steps: trigger detection and argument detection. For example, in ‘[the patient] was put on [Imatinib regimen]’, the first step detects the event trigger ‘put’, and the second step detects the theme ‘Imatinib regimen’ and target argument ‘patient’. Björne et al. [64, 66] converted sentences to a dependency graph (Stanford Dependency [111]) using the McClosky-Charniak-Johnson parser [112, 113] and explored the graphs to construct features for both steps. The McClosky-Charniak-Johnson parser is based on the constituency parser of Charniak and Johnson [112] and retrained with the biomedical domain model of McClosky [113]. Björne et al. generated n-gram features connecting event arguments based on the shortest path of syntactic dependencies (also see [114] for their original use in kernel method) between the arguments. They included as features the types and super-types of trigger nodes from event-type generalization, to address feature sparsity. They also applied semantic post-processing to prune graph edges that violate semantic compatibility as required by the event definition to hold between event arguments. Their system TEES performed best in the 2009 GE (0.52 f-measure), 2011 EPI (0.5333 f-measure), 2013 CG (0.5541 f-measure), 2013 GRO (0.215 f-measure, being the only participating system) and 2013 BB full-event extraction (0.14 f-measure). Hakala et al. [67] built on top of the TEES system and re-ranked its output by enriched graph-based features, including paths connecting nested events and occurrence of gene–protein pairs in general subgraphs mined from external PubMed abstracts and the PubMed Central full-text corpus. The system by Hakala et al. placed first in 2013 GE (0.5097 f-measure), whereas the TEES system placed second (0.5074 f-measure). The strong performance of both systems with heavy utilization of graph-based features, especially the fact that Hakala et al. extended Björne et al. using enriched graph-based features and obtained better performance (first versus second place), suggests the potential benefits of exploring graph-based features.

Miwa et al. [80, 81] built the EventMine system that can extract not only biomedical events but also their negations and uncertainty statements. For event extraction, they used the Enju parser [115] and the GENIA Dependency parser (GDep) [116] to generate path features along with dictionary-based features (e.g. UMLS Specialist lexicon [117] and Wordnet [118]). Their entry in BioNLP-ST-2013 placed first in the PC task. In particular, their path features include not only paths between event arguments but also paths between event argument and non-argument named entities. The latter paths likely account for the strong performance by providing more local context features.
Table 1. Summarization and characterization of relation extraction systems

<table>
<thead>
<tr>
<th>Methods</th>
<th>Parsers(^a)</th>
<th>Graph algorithm and usage(^b)</th>
<th>Information</th>
<th>CoRef</th>
<th>External resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. [60, 61], Mackinlay et al. [62], Ravikumar et al. [63]</td>
<td>McCJ, SD</td>
<td>Exact subgraph matching, approximate subgraph matching to match dependency graphs to event graphs</td>
<td>TN, DE</td>
<td>No</td>
<td>PDB, Uniprot, Biothesaurus</td>
</tr>
<tr>
<td>Björne et al. [64, 66], Hakala et al. [67]</td>
<td>McCJ, SD</td>
<td>Identifying shortest path to generate features for event arguments, rule-based graph pruning to exclude invalid event arguments</td>
<td>NEN, REN, EAE</td>
<td>No</td>
<td>Uniprot, SubtiWiki, Wordnet, DrugBank</td>
</tr>
<tr>
<td>Kilicoglu et al. [68, 69]</td>
<td>McCJ, SD</td>
<td>Transforming dependency graphs to embedding graphs, and rule-based traversal of embedding graphs to map them to events</td>
<td>NEN, REN, DE</td>
<td>Yes</td>
<td>Compiled dictionaries</td>
</tr>
<tr>
<td>Hakenberg et al. [70, 71]</td>
<td>BioLG</td>
<td>Using customized query language and post-processing rules to match subgraph patterns to events</td>
<td>TN, DE</td>
<td>Yes</td>
<td>Compiled dictionaries, Uniprot, GO</td>
</tr>
<tr>
<td>Riedel et al. [72, 73]</td>
<td>McCJ</td>
<td>Scoring candidate graphs to rank the event arguments</td>
<td>NEN, EAE</td>
<td>No</td>
<td>Compiled dictionaries</td>
</tr>
<tr>
<td>Van Landeghem et al. [74]</td>
<td>Stanford</td>
<td>Identifying minimal event containing subgraph patterns to construct event extraction rules</td>
<td>TN, DE</td>
<td>No</td>
<td>Compiled dictionaries</td>
</tr>
<tr>
<td>Kaljurand et al. [75]</td>
<td>Pro3Gres</td>
<td>Enumerating dependency paths between the concept pairs to count frequencies of paths in events and calculate likelihood of event arguments</td>
<td>NEN, DE</td>
<td>Yes(^1)</td>
<td>IntAct</td>
</tr>
<tr>
<td>Vlachos et al. [76]</td>
<td>RASP</td>
<td>Enumerating dependency paths between the concept pairs to manually identify paths that are likely to generate events</td>
<td>TN, DE</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>McClosky et al. [77, 78]</td>
<td>McCJ, SD</td>
<td>Using minimum spanning tree algorithm to parse events directly from sentences</td>
<td>NEN, EAE</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Quirk et al. [79]</td>
<td>McCJ, SD; Enju</td>
<td>Identifying shortest paths between the concept pairs to generate dependency chain as features</td>
<td>TN, DE</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Miwa et al. [80, 81]</td>
<td>Enju, GDep</td>
<td>Enumerating dependency paths between the concept pairs to generate dependencies chain as features</td>
<td>TN, DE</td>
<td>No</td>
<td>UMLS, Wordnet</td>
</tr>
<tr>
<td>Coulet et al. [40, 82], Percha et al. [42]</td>
<td>Stanford</td>
<td>Enumerating dependency paths between a named entity and a verb, then merging paths ending in the same verb to form binary relations</td>
<td>TN, DE</td>
<td>No</td>
<td>PharmGKB</td>
</tr>
<tr>
<td>Hakenberg et al. [83]</td>
<td>Stanford</td>
<td>Enumerating subtrees rooted at the lowest common ancestors of candidate concept pairs to subsequently pick the closest pairs</td>
<td>TN, CE</td>
<td>No</td>
<td>UMLS, SIDER, DrugBank, PharmGKB, GNAT</td>
</tr>
<tr>
<td>Wang et al. [84]</td>
<td>No</td>
<td>Assigning KL divergence as distance of an edge in a association graph, identifying an entity with shortest weighted distance path from</td>
<td>Nodes and edges from a curated association graph, AE</td>
<td>No</td>
<td>Chem2Bio2RDF</td>
</tr>
</tbody>
</table>

\(^a\) | \(^b\)
<table>
<thead>
<tr>
<th>Methods</th>
<th>Parsers</th>
<th>Graph algorithm and usage</th>
<th>Information</th>
<th>CoRef</th>
<th>External resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bui et al. [85]</td>
<td>Stanford</td>
<td>Using grammatical rules to traverse the tree structures to extract relations</td>
<td>TN, CE</td>
<td>Yes</td>
<td>HIVDB, RegaDB</td>
</tr>
<tr>
<td>Katrenko et al. [86]</td>
<td>LGP, Minipar, Charniak</td>
<td>Enumerating the lowest common ancestors of concept pairs in the dependency tree to be used as features</td>
<td>TN, DE</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Sætre et al. [87]</td>
<td>Enju, GDep</td>
<td>Enumerating dependency paths between the concept pairs to generate dependency chain as features</td>
<td>NEN, DE</td>
<td>No</td>
<td>UniProt, Entrez Gene, GENA</td>
</tr>
<tr>
<td>Weng et al. [36]</td>
<td>In-house parser</td>
<td>Frequent subtree pattern mining to construct relations representing clinical research eligibility criteria</td>
<td>SNEN, DE</td>
<td>No</td>
<td>UMLS</td>
</tr>
<tr>
<td>Thomas et al. [88]</td>
<td>McCJ, SD</td>
<td>Using Graph kernels (e.g. APG, kBSPS) to compare relation instance similarity</td>
<td>TN, DE</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Chowdhury et al [89–91]</td>
<td>Stanford, McCJ, SD</td>
<td>Using tree kernel MEDT to compare relation instance similarity</td>
<td>TN, DE</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Luo et al. [49, 50, 92]</td>
<td>Stanford (augmented by UMLS)</td>
<td>Frequent subgraph mining with redundancy removing to extract complex relations without supervision</td>
<td>SNEN, DE</td>
<td>No</td>
<td>UMLS</td>
</tr>
<tr>
<td>Roberts et al. [93]</td>
<td>Stanford</td>
<td>Identifying shortest paths between the concept pairs to compute edit distances between relation instances</td>
<td>TN, DE</td>
<td>No</td>
<td>UMLS, Wordnet, Wikipedia</td>
</tr>
<tr>
<td>deBruijn et al. [94]</td>
<td>McCJ, SD</td>
<td>Identifying minimal trees over concept pairs to generate dependency paths between concept covering minimal trees as features</td>
<td>TN, DE</td>
<td>No</td>
<td>UMLS</td>
</tr>
<tr>
<td>Xu et al. [95]</td>
<td>Kay</td>
<td>Customized parsing to produce conceptual graph representation</td>
<td>NEN, EAE</td>
<td>No</td>
<td>UMLS</td>
</tr>
<tr>
<td>Solt et al. [96]</td>
<td>Stanford, McCJ, Enju</td>
<td>Using graph kernels (e.g. APG, kBSPS) to compare relation instance similarity</td>
<td>TN, DE</td>
<td>No</td>
<td>Compiled dictionaries</td>
</tr>
<tr>
<td>Xu et al. [97]</td>
<td>Stanford</td>
<td>Extracting dependencies (edges) directly as features to classify relations; using CRF to classify body location relation</td>
<td>TN, DE (PE for CRF)</td>
<td>No</td>
<td>UMLS</td>
</tr>
<tr>
<td>Pathak et al. [98]</td>
<td>No</td>
<td>Using dictionary augmented CRF to detect relation keywords</td>
<td>TN, PE</td>
<td>No</td>
<td>Compiled dictionaries, UMLS</td>
</tr>
</tbody>
</table>

Abbreviation used in this table include: CoRef, coreference resolution; CRF, conditional random field; HMM, hidden Markov model; APG, all paths graph kernel [20]; kBSPS, k-band shortest path spectrum kernel [99]; MEDT, mildly extended dependency tree kernel [100]; PDB, Protein Data Bank [101]; UMLS, Unified Medical Language System. The key for the parsers are: Stanford—Stanford Parser, McCCJ—McClosky-Charniak-Johnson Parser, Chart—Kay Chart Parser, Enju—Enju Parser, Bikel—Bikel Parser, SD—Stanford Dependency. The key for the information represented by graphs are: TN—token as node, NEN—named entity as node, SNEN—semantically labeled named entity as node, REN—relation/event as node, AE—association as edge, PE—position adjacency as edge (e.g. connecting adjacent tokens), DE—dependency as edge, CE—constituency as edge, EAE—event argument as edge. When Stanford Parser is used with dependency as edge, Stanford Dependency is assumed by default.

*Parsers column is filled for only those systems that use dependency or constituency information from the parses.

*Graph Algorithm and Usage column is filled using the format [graph algorithm] to [usage].
Another vein of work proposed joint event extraction in which triggers and arguments for all events in the same sentence are predicted simultaneously. McClosky et al. [77, 78] integrated event extraction into the overall dependency parsing objective. They applied the McClosky-Charniak-Johnson parser and converted the parses to Stanford Dependency. They converted the annotated event structures in the training data to event dependency graphs that take event arguments (named entities) as nodes and argument slot names as edge labels. They mapped the event dependency graphs to Stanford Dependency graphs and generated graph-based features to train an extended MSTParser [119] for extracting event dependency graphs from test data. Their graph-based features included paths between nodes in the Stanford Dependency graph, as well as subgraphs consisting of parents, children and siblings of the path nodes. They then converted the top-n extracted event dependency graphs back to event structures and re-ranked event structures for the best one, using the same graph-based features from event dependency graphs. Riedel et al. [120] first applied Markov Logic Networks to learn event structures and later switched to graph-based methods [72, 73]. They projected events to labeled graphs, and scored candidate graphs using a function that captures constraints on event triggers and arguments. The scoring function considers token features, dictionary features and dependency path features. Riedel et al. further used a stacking model to combine their system with the system by McClosky et al. [77, 78]. The combined system obtained first place in 2011 GE task (0.56 f-measure) and 2011 ID task (0.556 f-measure).

Most of the remaining BioNLP systems that performed competitively also used graph-based features to various extents. Liu et al. [60] developed an Exact Subgraph Matching (ESM) method, and later a more flexible Approximate Subgraph Matching (ASM) method [61–63]. They processed sentences with the McClosky-Charniak-Johnson parser and transformed the parses to directed dependency graphs. They constructed the graph of an event by computing unions of dependency paths between event arguments. Liu et al. then applied ESM/ASM from sentence graphs to event graphs, using a customized distance metric that accounts for subgraph differences in graph structure, node-covered stemmed words and edge directionality. They extended their method by integrating Protein Data Bank (PDB) [101], Uniprot [121] and Biothesaurus [122] to recognize protein/residue names and also adapted their method to other corpora including BioInfer [123] and Uniprot corpus [124]. This work falls along the lines of graph kernel-based methods. In general, kernel methods weight similarity function explicitly, and features are only used to evaluate similarity function on instance pairs. Thus kernel methods usually cannot directly weight/rank features. Kilicoglu et al. [68, 69] also adopted the McClosky-Charniak-Johnson Parser/Stanford Dependency pipeline. They converted the dependency graphs to embedding graphs, where nodes themselves can be small dependency graphs, to apply post-processing rules to traverse embedding graphs and extract nested events. However, their embedding graphs also lead to argument error propagation and hurt precision.

Besides the popular McClosky-Charniak-Johnson Parser/Stanford Dependency pipeline, some systems experimented with different parsers and/or dependency representations. Hakenberg et al. [70] applied BioLG [125], a Link Grammar Parser [126] extension, to generate parse trees. They stored parse trees in a database and designed a query language to match subgraph patterns, which are manually generated from training data, against parse trees. They pointed out that generalization of event types could further improve their results. Van Landeghem et al. [74] analyzed dependency graphs from the Stanford Parser [127], identified minimal event-containing subgraph patterns from training data and constructed extraction rules based on these patterns. Their post-processing rules handled overlapping triggers of different event types and events based on the same trigger, trading recall for precision.

The other systems generally used the dependency paths connecting the concept pairs as features for event extraction. For example, the dependencies were obtained through applications of different parsers including the Pro3Gres parser [128] (Kaljurand et al. [75]), the RASP parser [129] (Vlachos et al. [76]) or both McClosky-Charniak-Johnson parser and Enju parser (Quirk et al. [79]). However, most of these methods attained inferior performance compared with the best systems in the same tasks. We believe that there are at least two reasons: the McClosky-Charniak-Johnson parser with the self-trained biomedical parsing model is probably the most accurate parser in this domain; the enriched graph-based features and event type generalization as used by the best systems likely produced more useful features for event extraction.

**PPI extraction and BioCreative shared tasks**

BioCreative shared tasks focused on automatic named entity recognition on genes and proteins in biomedical text and on extraction of the interactions between these entities [130–133]. Among the participants of the PPI task of BioCreative-II [132], most systems used co-occurrence statistics, pattern templates and shallow linguistic features (e.g. context words and part-of-speech tags), with either statistical machine learning or rule-based systems. Some systems observed the need for capturing cross sentence mentions of interacting proteins. For example, Huang et al. [134] developed a profile-based method that creates a vector representation for candidate protein pairs by aggregating features from multiple sentences in the document.

---

**Table 2. BioNLP event extraction tasks from the 2009, 2011 and 2013 shared tasks**

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Task descriptions</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE</td>
<td>Extracting the biomolecular events related to NF, NFB proteins.</td>
<td>2009, 2011, 2013</td>
</tr>
<tr>
<td>EPI</td>
<td>Extracting epigenetic and post-transcriptional modification events.</td>
<td>2011</td>
</tr>
<tr>
<td>ID</td>
<td>Extracting events describing the biomolecular foundations of infectious diseases.</td>
<td>2011</td>
</tr>
<tr>
<td>BB</td>
<td>Extracting the association between bacteria and their habitats.</td>
<td>2011, 2013</td>
</tr>
<tr>
<td>BI</td>
<td>Extracting the bacterial molecular interactions and transcriptional regulations.</td>
<td>2011</td>
</tr>
<tr>
<td>CG</td>
<td>Extracting cancer-related molecular- and cellular-level foundations, tissue- and organ-level effects and organism-level outcomes.</td>
<td>2013</td>
</tr>
<tr>
<td>PC</td>
<td>Extracting signaling and metabolic pathway-related biomolecular reactions.</td>
<td>2013</td>
</tr>
<tr>
<td>GRO</td>
<td>Extracting regulatory events between genes.</td>
<td>2013</td>
</tr>
</tbody>
</table>

---

1 Relative clause anaphora
In BioCreative-II.5, based on the top teams in the PPI task, the organizers pointed out that the BioNLP techniques using deep parsing and dependency tree/graph mining were necessary to achieve significant results [133]. In particular, Hakenberg et al. [71] used a system similar to their BioNLP-ST-2009 entry system [70]. They manually generated subgraph patterns from training data and matched them against parse trees. Their f-measure of 0.30. Sætre et al. [87] applied the Enju parser and the GDep parser and considered the dependency paths between concept pairs as features for relation extraction. They achieved an f-measure of 0.374. The PPI tasks of BioCreative-III consisted of detecting PPI-related articles that provide evidence to specified PPIs, but did not include the actual extraction of PPIs, which is the focus of this review [131]. Several follow-up studies to BioCreative-II.5 evaluated the usage of kernels in PPI extraction [99, 135], based on corpora including AlMed [136], BioInfer [123], HPReS [137], IEPA [138] and LLL [139]. They categorized kernels into the following categories: (1) shallow linguistic (SL) kernels [140]; (2) constituent parse tree-based kernels, including subtree (ST) [141], subset tree (SST) [142] and partial tree (PT) [143] kernels that use increasingly generalized forms of subtrees, as well as a spectrum tree (SpT) [144] kernel that uses path structures from constituent parse trees; (3) dependency parse tree-based kernels, including edit distance and cosine similarity kernels that are based on shortest paths [145], k-band shortest path spectrum (kBSPS) [99] that additionally allows k-band extension of shortest paths, all-path graph (APG) kernel [20] that differently weights shortest paths and extension paths in similarity calculation, as well as Kim’s kernels [146] that combine lexical, part-of-speech and syntactic information along with the shortest path structures. The comparative studies and error analyses showed that: (1) dependency tree-based kernels generally outperform constituent tree-based kernels; (2) kernel method performances heavily depend on corpus-specific parameter optimization; (3) APG, kBSPS and SL are top-performing kernels; (4) ensembles based on dissimilar kernels can significantly improve performance; (5) non-kernel-based methods (e.g., rule-based method, BayesNet) can perform on par with or better than all-top kernel methods. From these observations, it is evident that richer dependency graph/tree structures (e.g., in APG, kBSPS) than shortest paths are important to better performance of graph/tree-based kernels, which is consistent with the analysis of BioNLP participating systems. The limited advantage of the kernel methods over non-kernel methods and the interpretation difficulty associated with kernel methods seem to favor investigating novel feature sets rather than novel kernel functions.

DDI extraction and DDIExtraction shared tasks

The two DDIExtraction challenges (2011 and 2013) aimed at automated extraction of DDI from biomedical text [55, 56]. The organizers of the two challenges recognized the extended delays in updating manually curated DDI databases. They observed that the medical literature and technical reports are the most effective sources for the detection of DDIs but contain an overwhelming amount of data. Thus, DDIExtraction was motivated by the pressing need for accurate automated text-mining approaches. DDIExtraction-2011 focused on classifying whether there is any interaction between candidate drug pairs. DDIExtraction-2013 additionally pursued detailed classification of DDIs into one of the four possible subtypes: advice (advice regarding the concomitant use of two drugs), effect (effect of DDI), mechanism (pharmacodynamics or pharmacokinetic mechanism of DDI) and int (general mention of interaction without further detail). In DDIExtraction-2011, Thomas et al. [88] applied the McClosky-Charniak-Johnson parser/Stanford dependency pipeline. They used voting to combine the following kernels to implicitly capture features for relation extraction: APG [20], kBSPS [99] and SL [140] kernels. Their system achieved the best f-measure of 0.657. Chowdhury et al. [89, 90] applied the Stanford parser to obtain dependency trees and experimented with both feature-based methods and kernel-based ensemble methods for relation extraction. They experimented with SL [140], mildly extended dependency tree [100] (expanding shortest paths to also cover important verbs, modifiers or subjects) and path-encoded tree [147] (based on constituency tree) kernels. By combining feature-based and kernel-based methods, Chowdhury et al. achieved the second best f-measure of 0.6398.

In DDIExtraction-2013, Chowdhury et al. [89, 90] used their previous kernel method but switched to the McClosky-Charniak-Johnson parser and converted the parses to Stanford Dependency [91]. They attained an f-measure of 0.80 for general classification and 0.65 for detailed classification and placed first. Thomas et al. [148] followed a two-step approach to first detect general DDIs and then classify detected DDIs into subtypes. For the general DDI task, they used voting to combine kernels including APG [20], subtree (ST) [141], SST [142], SpT [144] and SL [140] kernels. For the subtype classification step, they used TEES directly [66]. Their system performed second best with an f-measure of 0.76 for general classification and 0.609 in detailed classification. It is interesting to see that adoption of systems originally developed for PPI extraction or event extraction has led to top performances in DDIExtraction. This further corroborates that these tasks are closely related, and technical solutions for one are generalizable to others.

Pharmacogenomics

In pharmacogenomics, numerous efforts have centered on the utilization of literature and clinical text to mine interesting relations between genetic mutations and drug response phenotypes. Although it is difficult to compare their performances because the experiments are not on shared corpora, these approaches do illuminate the translational application and adaptation of some state-of-the-art biomedical relation extraction techniques to problems directly asked by clinicians and pharmacologists.

Some systems used path-based approaches. Coulet et al. [40, 82] aimed at extracting binary relations between genes, drugs and phenotypes to build semantic networks for pharmacogenomics. They first converted the Stanford Parser output on sentences (from collected PubMed abstracts) into dependency graphs. They tracked the paths starting from named entities and ending at a verb, and merged paths ending with the same verb to form binary relations. Coulet et al. retained frequent relations and normalized both the collected entities and relation types (verbs). Without requiring prior enumeration of relation types, they created a semantic network knowledge base from 17 million MEDLINE abstracts, providing semantically rich summaries of pharmacogenomics relations. Percha et al. [42] extended this approach to use breadth-first search to yield the shortest path between two named entities in the dependency graph to generate features for relation extraction. They combined the extracted gene–drug relations to infer DDIs for those drugs that interact with the same gene product. Wang et al. [84] used Latent Dirichlet Allocation to create a semantic representation of biomedical named entities and used Kullback-Leibler (KL)
divergence to calculate the association distance between pairs of entities in the Chem2Bio2RDF [149] semantic network. They ranked candidate associations between named entity pairs by summing distances along the path connecting the pairs. They demonstrated uses cases on novel knowledge discovery including searching and predicting novel gene–drug associations, traversing the mined semantic network to compare the molecular therapeutic and toxicological profiles of candidate drugs.

Other systems used tree-based approaches. Katrenko et al. [86] studied gene–protein relation and protein–protein relation extraction and included as features the subtrees rooted at the lowest common ancestors of two named entities in the dependency parse trees. Their experiment used several parsers including the Link Grammar Parser [126], Minipar [150] and the Charniak Parser [151]. Compared with individual parser’s results separately, they reported improved performance from adopting ensemble methods (stacking and AdaBoost) and combining multiple parsers’ results [152]. Hakenberg et al. [83] aimed at extracting relations among genes, single-nucleotide polymorphism variants, drugs, ADRs. They relied on co-occurrence for extraction of certain relations (e.g. gene–drug, gene–disease and drug–disease), but augmented co-occurrence with subtrees from the Stanford Parser output for other types of relations. They considered binary relations and used subtrees rooted at the lowest common ancestors of named entity pairs. The mined relations are cross-referenced with knowledge bases including EntrezGene [153], PharmGKB [17] and PubChem [154]. Bui et al. [85] aimed at extracting relations between drugs and virus mutations from the literature to predict HIV drug resistance. They used Stanford Parser to generate constituent parse trees for sentences and developed grammatical rules that traverse the tree structures to extract drug–gene relations. Their system is in research use at five hospitals to preselect novel HIV drug resistance candidates.

Both path-based and tree-based systems in pharmacogenomics tend to focus on precision over recall in their evaluation, differing from the balanced f-measure used in multiple shared tasks. This likely stems from their specific goals of harvesting reliable relations to build and grow pharmacogenomics semantic networks. Too much noise will likely cloud the initial semantic network, while missing relations still have a chance to be later discovered with growing literature. In fact, reported precisions for pharmacogenomics relation extraction systems typically range from 70% to >80%. In addition, these systems often check extracted relations against curated database such as PharmGKB. We believe that these systems can further benefit from adopting parsers trained with biomedical models and using enriched graph-based features, two of the most recent lessons learned in shared tasks.

**Relation extraction from clinical narrative text**

The medical informatics community has also extensively studied relation extraction in the form of shared tasks and separately motivated research. For example, significant advances in extracting semantic relations from narrative text in EMRs have been documented in the i2b2/VA-2010 challenge (i2b2—Informatics for Integrating Biology to the Bedside, VA—Veterans Association) [1].

**i2b2/VA challenge**
The challenge had three tasks including concept extraction, assertion classification and relation classification, participated by numerous international teams [1]. Concept extraction can be considered the basic task, as assertions and relations all refer to the extracted concepts. As the challenge allows relation classification to use the ground truth of concepts extraction, the performance metrics for relation classification should be interpreted as an upper bound for the end-to-end relation extraction task (same as the challenges from BioNLP, BioCreative and DDIExtraction). In this section, we review only the relation classification systems, where the target relations are predefined among medical problems, tests and treatments. These relations include ‘treatment improves / worsens / causes / is administered for / is not administered because of medical problem’, ‘test reveals / conducted to investigate medical problem’ and ‘medical problem indicates medical problem’. As in reviewing the above challenges, we review only those systems that represented sentences as graphs and explored such graphs during the feature-generation step.

Roberts et al. [93] classified the semantic relations using a rather comprehensive set of features: context features (e.g. n-grams, GENIA part-of-speech tags surrounding medical concepts), nested relation features (relations in the text span between candidate pairs of concepts), single concept features (e.g. covered words and concept types), Wikipedia features (e.g. concepts matching Wikipedia titles), concept bi-grams features and similarity features. The latter were computed using edit distance on language constructs including GENIA phrase chunks and Stanford Dependency shortest paths. Their system reached the highest f-measure on relation classification (0.737).

deBrujin et al. [94] applied a maximum entropy classifier with down sampling applied to balance the relation distribution. They applied the McClosky-Charniak-Johnson parser/Stanford Dependency pipeline, and included as features the dependency paths between the minimal trees that cover the concept pairs. They used word clusters as features to address the problem of unseen words. Their system reached the second best f-measure of 0.731. Solt et al. [96] then experimented with several parsers including the Stanford Parser, the McClosky-Charniak-Johnson Parser and the Enju Parser. They used the resulting dependency graphs with two graph kernels including the all paths graph (APG) kernel [20] and kBSPS [99], which produced only moderate performance. This likely reflects the difficulty in tuning the graph/tree kernel-based systems, consistent with the observations from the experience in relation/event extraction from the scientific literature.

**SemEval 2015 Task 14**
The SemEval 2015 Task 14 included disorder identification and disorder slot filling tasks [155]. Disorder identification is essentially named entity detection, and disorder slot filling is similar to BioNLP event extraction tasks but in clinical subdomain. The challenge further divided the slot filling task into two subtasks, one with gold-standard disorder spans (task 2a) and one without (task 2b). Thus, task 2b has stricter evaluation results than task 2a. The attribute slots defined by the challenge include concept unique identifier (CUI), negation (NEG), subject (SUB), uncertainty (UNC), course (COU), severity (SEV), conditional (CND), generic (GEN) and body location (BL). Identifying the CUI is the named entity-detection problem, and identifying negation and uncertainty is the assertion classification problem. Identifying SUB, COU, SEV, CND, GEN and BL are more analogous to binary relation extraction. They are not completely equivalent to binary relation extraction, as the challenge limited the possible values for those slots, adding a layer of abstraction.

The challenge used weighted accuracy to rank the participants. Xu et al. [97] and Pathak et al. [98] consistently ranked as
the top two teams in both task 2a (0.886 and 0.880, respectively) and task 2b (0.808 and 0.795, respectively). Xu et al. used Conditional Random Field (CRF) as the classifier for BL slot filling and SVM as the classifier for the other slots. The SVM classifier additionally used dependencies coming into and out of the disorder mentions. Note that these dependencies cannot capture multi-hop syntax dependence, but the authors observed that NEG/UNC/COU/SEV/GEN always have one-hop dependence. On the other hand, CRF (for BL) is itself a graph-based model that treats tokens and hidden states as nodes (integrating semantic and syntactic features including n-grams, context words, dictionaries and section names) and interconnects nodes with transition and emission edges [156]. Pathak et al. divided slot detection into two parts: detecting keywords and relating keywords with disorder mentions. They used dictionary look-up combined with CRF trained on features such as bag-of-words and orthographic features to detect keywords. To relate keyword with disorder mentions, they trained SVM using features similar to Xu et al. plus Part-of-Speech tags. Other teams used explicit graph-mining algorithms [157, 159] and did not perform as competitively. For example, Hakala et al. [157] tackled task 2a by adapting TESS system to work with SemEval data format and achieved a weighted accuracy of 0.857, placing the third. This is not surprising, as given many slots only involve one-hop dependencies, full-fledged graph-based approach only offers limited benefits. In addition, the controlled vocabulary and controlled format nature of challenge tasks makes themselves suitable for CRF, as limited number of states and state-transitions lead to less sparse and more robust probability estimation.

Separately motivated clinical relation extraction

After the i2b2 challenges, several authors aimed at combining the concept and relation extraction steps into an integral pipeline and/or generalizing to the extraction of complex or even nested relations. Xu et al. [95] developed a rule-based system MedEx to extract medications and specific relations between medications and their associated strengths, routes and frequencies. The MedEx system converts narrative sentences in clinical notes into conceptual graph representations of medication relations. To do so, Xu et al. designed a semantic grammar directly mappable to conceptual graphs and applied the Kay Chart Parser [160] to parse sentences according to this grammar. They also used a regular-expression-based chunker to capture medications missed by the Kay Chart Parser. Weng et al. [36] applied a customized syntactic parser on text specifying clinical eligibility criteria. They mined maximal frequent subtree patterns and manually aggregated and enriched them with the UMLS to form a semantic representation for eligibility criteria, which aims to enable semantically meaningful search queries over ClinicalTrials.gov. Luo et al. [49] augmented the Stanford Parser with UMLS-based concept recognition to accurately generate graph representations for sentences in pathology reports where the graph nodes correspond to medical concepts. Frequent subgraph mining was then used to collect important semantic relations between medical concepts (e.g. which antigens are expressed on neoplastic lymphoid cells), which serve as the basis for classifying lymphoma subtypes. Extending the subgraph-based feature generation into unsupervised learning, Luo et al. [50] further used tensor factorization to group subgraphs. The intuition is that each subgraph corresponds to a test result, and a subgraph group represents a panel of test results, as typically used in diagnostic guidelines. The tensors incorporated three dimensions: patients, common subgraphs and individual words in each report. The words helped better group subgraphs to recover lymphoma subtype diagnostic criteria.

Shared resources for relation extraction

The shared tasks and separately motivated research on biomedical relation extraction have not only advanced the state-of-the-art in methodology, but also created and/or demonstrated the utilization of a repository of shared resources that range from knowledge bases to shared corpora to graph mining toolkits. We categorize and summarize those resources in Table 3, in the hope that it may serve as a starting point for resource navigation for future research efforts. Some of those resources concern general domain (e.g. general terminology/ontology resources Wordnet [118], Verbnet [163]); some concern the biomedical domain comprehensively (e.g. domain-specific terminology/ontology resources Gene Ontology [161], UMLS [58], Medical Subject Heading [162] and Biothesaurus [122]); some target-specific biomedical subdomain (e.g. knowledge bases such as PDB [101], Uniprot [121], SIDER [164], DrugBank [165], HIVDB [166], RegaDB [167], Entrez Gene [153], GENA [18] and IntAct [169]).

The road ahead

Although notable progress has taken place in applying graph-based algorithms to improve the extraction of biomedical relations, barriers still exist to developing practical relation extraction methods that are both generalizable and sufficiently accurate. Below we discuss a few such barriers and potential directions to overcome them.

Not all parsers and dependency encodings are synergistic

It has been pointed out repeatedly that the choice of the parser and dependency encodings may play an important role in a relation extraction system’s performance. Buyko et al. [177] performed comparative analysis on the impact of graph encoding based on different parsers (Charniak-Johnson [112], McClosky-Charniak-Johnson, Bikel [175], GDep, MST [119], MALAT [178]) and dependency representations (Stanford Dependency and CoNLL dependency) and found that the CoNLL dependency representation performs better in combination with four parsers than the Stanford Dependency representation; and McClosky-Charniak-Johnson parser frequently places as the best performing parser. Miwa et al. [179] compared five syntactic parsers for BioNLP-ST-2009. They concluded that although performances from using individual parsers (GDep, C&C [176], McClosky-Charniak-Johnson, Bikel, Enju) do not differ much, using an ensemble of parsers and different dependency representations (Stanford Dependency, CoNLL, Predicate Argument Structure) can improve the event extraction results. As Stanford Dependency is the most widely used dependency encoding, they also compared the performance of using different Stanford Dependency variants and found that basic dependency performs best if keeping types of dependency edges. On the other hand, if ignoring types of dependency edges, they found that the collapsed dependency variant performs best, which corroborates the finding by Luo et al. [49]. In [49], the task is extracting relations as features to classify lymphoma subtypes instead of classifying relations themselves as in supervised relation classification in the BioNLP-ST event extraction tasks. Thus, recall is favored in
the feature learning step, where ignoring types of dependencies helps to improve the coverage of subgraph patterns. The lessons learned in [49, 177, 179] seem to corroborate individual reports from top participants from challenges and are consistent with the popularity of parser-dependency choice from non-challenge applications in sections Pharmacogonomics and Separately motivated clinical relation extraction. In particular, we expect a good combination is to choose from either the McClosky-Charniak-Johnson parser or Stanford parser augmented with medical lexicon and choose from either CoNLL dependency or collapsed Stanford dependency.

### Integrating coreference resolution

Coreference occurs frequently in biomedical literature and clinical narrative text, arising from the use of pronouns, anaphora and varied entities for the same concepts. Care must be exercised to transfer the correct relation along the coreference chain. However, many of the reviewed approaches for relation and event extraction did not have a built-in coreference resolution component. Miwa et al. [180] specifically studied the impact of using a coreference resolution system and showed improved event extraction performance. In particular, they developed a rule-based coreference resolution system that consists of detecting rules for mention, antecedent and coreferential link. They used the coreference information to modify syntactic parse results so that antecedent and mention share dependencies. Features were also extended between mentions and antecedents. Recognizing the importance of coreference features, several systems [181–184] subsequently implemented coreference resolution across sentences. They showed that to facilitate extracting specific types of relations, heuristic/rule-based coreference resolution tends to outperform domain-adapted statistical machine learning systems. Part of their lessons concerns the lack of gold-standard coreference annotation in relation extraction corpora; thus, we expect that paired relation and coreference annotations can improve coreference resolution and ultimately relation extraction. Moreover, integrating coreference resolution into the learning objective of relation extraction may lead to more coherent optimization and better end-to-end performance.

### General relation and event extraction and domain adaptation

The state-of-the-art relation and event extraction systems from shared tasks are mostly built around domain-specific definitions of relations and events, many of which are in fact binary (e.g. BioCreative PPI challenge [133], DDIExtraction challenge [55, 56] and I2b2/VA challenge [1]). However, there is a gap between the technical advances and the demands from many real-world tasks, including building pharmacogenomics

---

**Table 3. Shared resources for biomedical relation extraction**

<table>
<thead>
<tr>
<th>Utility category</th>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminology/Ontology/Knowledge Base</td>
<td>GO [161], UMLS [58], MeSH [162], Wordnet [118], Verbinet [163], Biocurators [122], PDB [101], Uniprot [121], SIDER [164], DrugBank [165], HIVDB [166], RegaDB [167], Entrez Gene [153], GENA [168], IntAct [169], Chem2Bio2RDF [149], PharmGKB [17], PubChem [154], SubtiWiki (<a href="http://subtiwiki.uni-goettingen.de/">http://subtiwiki.uni-goettingen.de/</a>)</td>
</tr>
<tr>
<td>Graph Miner</td>
<td>Gaston [170], Mofa [171], GSpan [172], FFSM [173], Graph Spider [174]</td>
</tr>
<tr>
<td>Tree/Graph Kernel</td>
<td>subtree (ST) kernel [141], subset tree (SST) kernel [142], partial tree (PT) kernel [143], spectrum tree (SpT) kernel [144], mildly extended dependency tree (MEDT) kernel [100], all-path graph (APG) kernel [20], k-band shortest path spectrum (kBSPS) kernel [99], path-encoded tree (PET) kernel [147]</td>
</tr>
<tr>
<td>Dependency Parsers</td>
<td>Enju Parser [115], GDep Parser [116], Stanford Parser [127], McCJ Parser [112, 113], RASP Parser [129], Bikkel Parser [175], BioLG Parser [125], Pro3GreParser [128], Kay Chart Parser [160], C&amp;C [176]</td>
</tr>
<tr>
<td>Shared Corpora</td>
<td>BioNLP-09 event corpus [108], BioNLP-11 event corpus [109], BioNLP-13 event corpus [110], BioCreative-II relation corpus [130], BioCreative-II.5 relation corpus [139], DDIExtraction relation corpora [55, 56], I2b2/VA corpus [1], i2b2/VA corpus [1], i2b2/VA corpus [1], i2b2/VA corpus [1], i2b2/VA corpus [1]</td>
</tr>
</tbody>
</table>

The resources are organized by their utility categories. Abbreviations used include: Gene Ontology (GO), Unified Medical Language System (UMLS), Medical Subject Heading (MeSH), Human Protein Reference Database (HPRD), Linguistic Data Consortium (LDC). Existing repositories are those that themselves list biomedical NLP resources.
semantic networks [40], extracting clinical trial eligibility criteria [36] and representing test results for automating diagnosis categorization [49, 50]. In those tasks, general relation and event discovery is necessary, where the number of nodes is flexible and even the relation/event structure may not be entirely prespecified. Systems exploring the automation of annotation scheme learning [66] or unsupervised relation extraction [49, 50] attempted to enable generalization but there is still large room for improving accuracy. Another challenge brought by domain-specific relation/event definition concerns the training data. The problem of limited training data often plagues the development of NLP systems, among which the relation extraction systems are no exceptions. To make better use of existing annotated corpora, it is necessary to perform domain adaptation from external training corpora (source) to the target corpora. Merely adapting top systems to the new format and training on the new corpora does not always lead to top performers in the new domains, e.g. the adapted TEES system [157] in SemEval 2015 Task 14. Miwa et al. [180] proposed adding source instances followed by instance reweighting when source and target match on events to be extracted. When source and target corpora have a partial match on events, they proposed to train each event extraction module separately on the source corpora and used its output as additional features for the corresponding modules on the target corpora. Miwa et al. [185] further improved methods of combining corpora by integrating heuristics to filter spurious negative examples. The heuristics aim to correct errors where instances with a ‘None’ annotation in one corpus owing to a different focus are all treated as negative instances in the combined corpus. Applying this method on learning from seven event-annotated corpora, they showed improved performance on two tasks in BioNLP-ST-2011. The successes from cascaded training [180] and example filtering heuristics [185] illuminate some promising directions in corpora adaptation. We expect parallel efforts in system generalization and corpora adaptation can complement each other toward effective domain adaptation.

Redundancy in subgraph patterns

Detection of useful subgraph patterns often depends on identifying ones with high frequency. A subgraph occurs once in a corpus whenever it is part of a larger graph in that corpus. Frequent subgraph mining identifies those subgraphs that occur in a corpus more than a given threshold number of times (see [186] for a survey on frequent subgraph mining and see [170–174] for various software). Lessons learned in multiple studies [49, 64, 65, 184] showed that redundancy among collected subgraph patterns creates problems when using subgraphs as features. Many smaller subgraphs are subisomorphic to (part of) other larger frequent subgraphs. Many of these larger subgraphs have the same frequency as their subisomorphic smaller subgraphs. This arises when a larger subgraph is frequent, and therefore all its subgraphs automatically become frequent as well. Furthermore, if the smaller subgraph $g_s$ is so unique that it is not subisomorphic to any other larger subgraph than $g_s$, then this pair $g_s, g_l$ shares identical frequency. Based on such observations, Luo et al. [49] only kept the larger subgraphs in such pairs. Note that it is cost prohibitive to perform a full pairwise check because the subisomorphism comparison between each subgraph pair is NP-complete [170], and a pairwise approach would ask for around a billion comparisons for a collection of several tens of thousands of subgraphs. Luo et al. presented an efficient algorithm using hierarchical hash partitioning that reduces the number of subgraph pairs to compare by several orders of magnitude. The key idea is that one only needs to compare subgraphs whose sizes differ by one, and one can further partition the subgraphs so that only those within the same partition need to be compared. The frequent subgraph redundancy is a systemic problem and is also closely related to the problem of nested terms in automatic term extraction, and a variety of de-duplicating scores (e.g. c-value [187]) proposed for the latter problem could be adapted to address the former problem. Other graph mining approaches may also be applicable, depending on the task. For example, algorithms may be developed to collect subgraph patterns that explore the ‘novelty’ of the subgraphs, such as using p-significance to assess how unusual it is to see the subgraphs in the current corpus [188].

Integrating with NER

Most shared task participants were not evaluated based on their relation extraction from scratch. Rather, their systems were evaluated given the gold standard of named entity annotations, which is even true for challenges that include a NER task, such as the i2b2/VA shared tasks. Thus, their evaluation results are likely an upper bound of the end-to-end system performance, the tuning of which is in fact a non-trivial task. Kabiljo et al. [189] evaluated several methods for end-to-end relation extraction including a keyword-based method, a co-occurrence-based method and a method using dependency-graph-based patterns. They noted that in general a significant performance drop will occur when using named entities tagged by NER system such as BANNER [190] instead of using the gold standard. In addition, it is useful but challenging to filter out unrelated named entity pairs, for which no relations have been explicitly stated in the text [92]. Such filtering may adopt a hybrid approach that relies on both automatically checking semantic type compatibility and manually sifting through the remaining tuples. However, as the number of non-related tuples often dominates that of related tuples, better automated filtering is necessary and remains an open question. Latest challenges such as SemEval 2015 Task 14 [155] added the end-to-end evaluation on joint detection/inference of entities and relations for prespecified relations, and it is reasonable to expect such evaluation will become more popular and will boost the development of end-to-end systems.

Relation extraction currently concentrates a large amount of efforts in biomedical information extraction. In the future, we anticipate a migration toward more unsupervised relation extraction that are increasingly adaptable across biomedical subdomains. The integration of relation extraction with named entity detection will produce end-to-end systems that can further automate the discovery and curation of novel biomedical knowledge. With advances in deriving better graph representations with more accurate parsers and appropriate dependency choices, in enhancing coreference resolution at document level, and in more efficiently sifting through informative subgraph patterns, the extraction of biomedical relations will continue to improve owing to an increase in the quality of data, and sustained community efforts. Given the rapid progress in the past years, we expect more exciting and promising developments of biomedical relation extraction, which continuously shape the emerging landscape and provide opportunities for researchers to contribute.
Key Points

- The state-of-the-art approaches on extracting biomedical relations from scientific literature and clinical narratives now heavily rely on graph-based algorithms to bridge semantic and syntactic information.
- State-of-the-art methods for biomedical relation extraction present multiple innovations from perspectives including strategies for parsing and dependency generation, graph exploration, as well as heuristics to address feature sparsity. They also create and maintain shared resources and tools for continuing research.
- Important applications of biomedical relation extraction include those in pharmacogenomics, clinical trial screening and adverse drug reaction detection, both advancing basic science and improving clinical practice.
- Challenges remain in biomedical relation extraction, including synergy between parsers and dependency encodings, integrating coreference resolution and named entity recognition, redundant subgraph patterns and domain adaptation. Insights from analyzing these challenges shed light on future directions to battle them.

Acknowledgement

The work was supported in part by Grant Number U54LM008748 from the National Library of Medicine, NIH 154HG007963 from the National Human Genome Research Institute and by the Scullen Center for Cancer Data Analysis.

References


