A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave

GIACOMO FRISONI[ID], GIANLUCA MORO[ID], AND ANTONELLA CARBONARO[ID]
Department of Computer Science and Engineering (DISI), University of Bologna, 47521 Cesena, Italy

Corresponding author: Giacomo Frisoni (giacomo.frisoni@unibo.it)

ABSTRACT **Motivation:** The scientific literature embeds an enormous amount of relational knowledge, encompassing interactions between biomedical entities, like proteins, drugs, and symptoms. To cope with the ever-increasing number of publications, researchers are experiencing a surge of interest in extracting valuable, structured, concise, and unambiguous information from plain texts. With the development of deep learning, the granularity of information extraction is evolving from entities and pairwise relations to events. Events can model complex interactions involving multiple participants having a specific semantic role, also handling nested and overlapping definitions. After being studied for years, automatic event extraction is on the road to significantly impact biology in a wide range of applications, from knowledge base enrichment to the formulation of new research hypotheses. **Results:** This paper provides a comprehensive and up-to-date survey on the link between event extraction and natural language understanding, focusing on the biomedical domain. First, we establish a flexible event definition, summarizing the terminological efforts conducted in various areas. Second, we present the event extraction task, the related challenges, and the available annotated corpora. Third, we deeply explore the most representative methods and present an analysis of the current state-of-the-art, accompanied by performance discussion. To help researchers navigate the avalanche of event extraction works, we provide a detailed taxonomy for classifying the contributions proposed by the community. Fourth, we compare solutions applied in biomedicine with those evaluated in other domains, identifying research opportunities and providing insights for strategies not yet explored. Finally, we discuss applications and our envisions about future perspectives, moving the needle on explainability and knowledge injection.

INDEX TERMS Biomedical text mining, event extraction, natural language understanding, semantic parsing.

I. INTRODUCTION

The scientific literature, which is the most important means of sharing new knowledge, has led to an information overload [1]. Biology is feeling most of this literature congestion blow, witnessing exponential growth in publications over the last few decades (Figure 1).

As of January 2021, PubMed alone has a total of 32 million articles dating back to the year 1781. Since 2011, more than 1 million new records pour into the PubMed database each year — about 2 papers per minute (became 3 in 2020).[1] Articles with valuable information on discoveries

The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi[ID].

[1]Bibliometric results obtained by executing the query ''1700:2021[dp]'' on the PubMed search engine, https://www.ncbi.nlm.nih.gov/pubmed/ (accessed Nov. 23, 2021).

and insights are continually released. To learn about the latest developments or to find relevant studies on a research field, scientists have to read more text documents in time than before. This task outstrips the processing capabilities of human domain experts, limiting the value of big data in biology and our ability to understand many diseases [2]. Consequently, there is an increasing demand for accurate biomedical text mining (TM) tools to automatically extract useful knowledge for each research community [3], ideally well organized and concise.

Deep learning has boosted the evolution of natural language processing (NLP), sparking cutting-edge advances in support of large-scale machine reading, biological analysis, and database curation. In the NLP field, information extraction (IE) is an effective approach to summarize knowledge into expressive forms for management and comprehension,
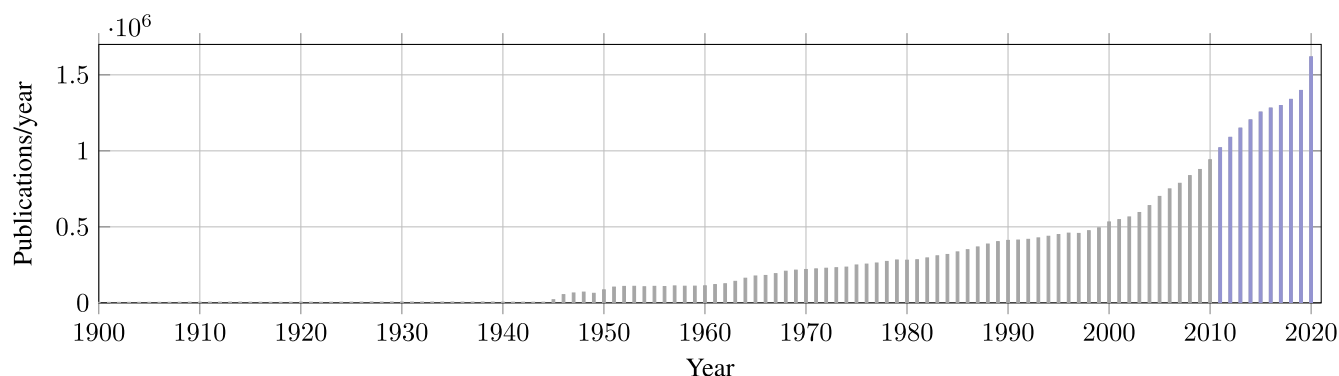
FIGURE 1. The annual rate of biomedical publications registered on PubMed, highlighting the years that have exceeded 1 million records.

thus emboldening decision making. Initial efforts in biomedical IE focused on two fundamental tasks: named entity recognition (NER) [4], [5] and named entity linking (NEL) [6]. The first aims to detect and classify mentions of entities in the text, such as names of genes, proteins, cellular components, or diseases. The second intends to link entity mentions to corresponding identifiers in reference knowledge bases (KBs), such as biology-related ontologies on BioPortal.[2] Although entity normalization still represents an open problem due to the ambiguity of language and the high-level branching of biomedical taxonomies,[3] the performance achieved by current research contributions is sufficient for numerous scenarios.

The real challenge that the biomedical community has been working on for several years to date is building accurate KBs of interactions between biologically relevant entities, possibly in a statistically quantifiable manner. To offer a quick glimpse at the problem, let's consider two simple sentences with opposite meaning: (a) *"protein X induces the expression of gene Y and gene Z"*, (b) *"protein X inhibits the expression of gene Y and gene Z"*. Mere entity co-occurrences do not necessarily indicate meaningful relationships and are semantically poor; i.e., $protein\_X \leftrightarrow gene\_Y \leftrightarrow gene\_Z$, without differentiating (a) from (b). Moreover, Chunk *et al.* [11] reported that only 30% of protein pairs co-occurring in the same sentences have an actual interconnection. To solve these issues, relation extraction (RE) has been researched for a long time, comprising drug-drug interactions (DDIs) [12], [13] and protein-protein interactions (PPIs) [13], [14]. RE systems can faintly discern (a) and (b) by hunting for precise predicates and expressing the direction of causality, but their targets are commonly limited to flat triplets like $protein\_X \xrightarrow{induces} gene\_Y$ and

$protein\_X \xrightarrow{inhibits} gene\_Y$ (with as many instances for $protein\_X$ and $gene\_Z$). As such, RE loses worthy details that seriously undermine its practical usefulness. Among them, sub-relationships (e.g., *expression_of*) and co-participation of more than two entities potentially covering separated functions (e.g., while $protein\_X$ is the regulation cause, $gene\_Y$ and $gene\_Z$ are the entities affected by the regulation).

On these premises, biomedical IE is notably shifting from the relational model to the event model, a more expressive method of capturing natural language statement semantics, enabling the formal representation of sophisticated biological processes. Event representations use a composite structure with multiple participants having a specific semantic role, including nested and overlapping definitions. This flexibility allows interconnecting entities with arbitrarily complex logic, a capability asked by many domains and downstream TM applications.

Since identifying events requires a deeper grasp of sentence meanings, event extraction (EE) is deemed relevant to natural language understanding (NLU). Automatically creating structured representations from large unstructured text corpora is helpful in a shedload of NLU tasks such as dialogue systems [15], question answering [16], and text summarization [17]. Ultimately, EE allows deriving exhaustive, interpretable, and quantifiable interactions otherwise buried in literature spans and not readily available for further usage. The latest NLP milestones are paving the way for totally new strategies and application opportunities, from the adoption of events for Explainable Artificial Intelligence (XAI) to the improvement of current language models through knowledge injection.

SUMMARY OF THE SURVEY

Some surveys have already dealt with the extraction of events from the text, maintaining a specific focus (decision support systems [18], [19], multilingual [20], social media [21], [22], biomedical [8], [23], [24], open domain [25]) or proposing a general and systematic review [26]. Even though the aforementioned articles testify to the upsurge of event-related activities, to our best knowledge, there is little work taking

---

[2]NCBO BioPortal. More than 800 ontologies and 10 million classes are currently available. https://bioportal.bioontology.org/ (accessed Nov. 23, 2021).

[3]Millions of biomedical entity names are already in use and constantly expanding, making consensus difficult to reach [7], [8]. The same words can be used to describe different bio-concepts; for example, "AP2" can be the name of a gene, a chemical, or a cell line. Conversely, the same bio-concepts can have different terms and span multiple words [9], [10].

into account the link between EE and NLU on a higher abstraction level, without adopting a closed view linked to the task or domain. We provide a broader view of events as semantic units and a rich literature taxonomy. In doing this, we focus on the biomedical domain for the recognized complexity of biomolecular events, the high number of contributions, and the potential impact deriving from their application. Our ultimate goal is to provide newcomers to the event NLP field with reference material that can serve as a new reading key and stimulate future research.

The remainder of this overview is organized as follows. Section II discusses the terminology and concepts revolving around events, ending up with a flexible definition and a structure/schema modeling distinction. Section III proceeds by thoroughly reviewing the EE task, its challenges and data sources, the proposed approaches and the current state-of-the-art (SOTA). Section IV reports observations on leading systems and examines the the research gap between biomedical and non-biomedical EE. Section V presents EE applications and future trends, with an outlook aimed at engaging the community around this vibrant research area, also reflecting on the benefits of structured event representations over alternative raw text-based solutions. Finally, Section VI concludes the survey.

II. EVENTS

Before proceeding with our literature examination, it is convenient to establish a common point of understanding on what the term *event* stands for in NLP and IE. We clarify the terminology used in the community and face thorny questions that remain open to this day [27], [28]. What is an event (and what is not)? What are the building blocks of an event? Does event knowledge have a common structure among all event types? In this section, we pause at the numerous definitions proposed by researchers to provide satisfactory answers.

A. DEFINITION

Events are a ubiquitous linguistic phenomenon. They originated from Cognitive Science [29] and have been extensively used by computational linguists in modeling the meanings of natural language expressions [30]. However, one of the issues that hinders establishing common grounds for event-based tasks is the absence of a unified definition of *event*, which usually varies in consonance with the research domain and purpose.

In Cognitive Science research, an event is often defined as *"something that happens or occurs"*, *"a spatiotemporal thing"*, *"an action"*, *"a unit for human beings to know the world"* [31], [32]. As recently stated in [33], *"a fundamental aspect of human cognition is the ability to parse our constantly unfolding experience into meaningful representations of dynamic events and to communicate about these events with others"*. From these observations, the key concept of *"dynamic knowledge"* emerges.

Most of the TM literature work refers to the definition proposed by the Automatic Content Extraction (ACE) research program, which we will further discuss in Section III-C. As stated in the ACE 2005 annotation guidelines [34], an event is *"a specific occurrence of something that happens in a certain time and a certain place involving one or more participants, which can frequently be described as a change of state"*. The view of real-world events as things that happen and linguistic events as their representation is very well established [35]. The ACE definition takes up this vision and focuses on the involvement of one or more entities (with status) linked together, forming a complete semantic unit. The notion of state change is crucial. In bioinformatics, an event generally refers to *"the state change of one or more biomedical entities (e.g., proteins, cells, chemicals) with interactions between them"* [8], [23], [36].

Nevertheless, the ACE definition is not sufficiently flexible as the temporal and spatial dimensions do not characterize all event types (like the biomedical ones), becoming auxiliary. In this vein, event-related information is commonly represented also in the "5W1H" form (i.e., who, when, where, what, why, and how), first introduced in 1949 [37] and broadly applied in journalism to judge article completeness. Hence, another frequent event definition is *"something done by somebody to someone else for some purpose, by some means, sometimes and somewhere"* [38]. Although the 5W1H approach is comparatively simpler than other more powerful meaning representations (e.g., semantic role labeling (SRL), abstract meaning representation (AMR)[4]) [39], the arguments' semantics do not necessarily respect this template. Consequently, it should only be seen as a general-domain indication.

To shed some light on this lack of consensus, bearing in mind the notes made above, we propose the following definition:

> *An event is a specific occurrence of something that happens and involves an arbitrary number of attributes and participants covering a specific semantic role, depending on the event type. The interaction (i.e., dynamic relation) modeled by an event represents or leads to some state change.*

In the biological field, it is critical to have detailed views on the behavior of biomolecules and their observable changes, which descriptions spread all over the literature [40]. For instance, biomolecular events englobe expression, transcription, localization, binding, phosphorylation, and regulation. Table 1 reports some practical examples.

Beyond the biomedical context, events allude to central objective information of the world, widely used to analyze news, social posts, and financial documents.

Discrete event representation (or event modeling) is a two-sided matter, made up of a structure and an optional schema. In Section II-B and Section II-C, we deepen these concepts, often confused and improperly interchanged.

---

[4]L. Banarescu *et al.*, "AMR Guidelines". https://github.com/amrisi/amr-guidelines (accessed Nov. 23, 2021).

TABLE 1. Sentences mentioning elementary biomedical events with annotated entities.

Event type(s)	Sentence
Binding	[Evi-1]$_{GENE}$ physically *interacts* with [Smad3]$_{GENE}$
Localization	[tumor tissue]$_{TISSUE}$ *liberates* [VEGF]$_{GENE}$ into the [blood]$_{ORGANISM\_SUBSTANCE}$
Regulation + Growth	We studied whether [L-NAME]$_{SIMPLE\_CHEMICAL}$ *affects* [adenocarcinoma]$_{CANCER}$ *growth*
Positive Reg. + Gene Expr.	[RFLAT-1]$_{PROTEIN}$ *activates* [RANTES]$_{GENE}$ *gene expression*

B. STRUCTURE

An event model is based on a specific semantic structure. Just as there is no standard agreement on event definition, the same is true for the event structure and its components. Many alternatives have been adopted in the field, proposed by organizations and individual researchers for distinct areas.

Again, the current prevailing structure is suggested by ACE (portrayed graphically in Figure 2). It stands on the subsequent key notions, which have evolved to manage increasingly complex events, where the meaning of "complex" will be clarified in Section III-A. These features usually adapt to any context, albeit textual and event characteristics can vary considerably between domains [41].

- **Event trigger**: textual mention (i.e., word or sequence of words) that more clearly expresses the occurrence of an event. It usually takes the form of verbs (e.g., "stimulates", "regulates") or normalized verbs (e.g., "expression", "regulation").
- **Event argument**: entity mention, value, or another event that serves as a participant or attribute in an event, fulfilling a specific role that characterizes its contribution. Entity mentions are typically proper nouns, such as "NF-kappa", "ATP", or "glucose". Attributes can be event-specific or general (e.g., time, place).
- **Argument role**: semantic relationship between an argument and an event. Typical examples of roles in biology are *Cause* (what is responsible for the event occurring) and *Theme* (what is affected by the event occurring).
- **Event property (or meta-knowledge [42])**: additional information useful for the interpretation of an event and specified in its textual context [24], [43]. Properties generally refer to event modifiers that reshape the described interaction (see Table 2); they may also apply simultaneously (e.g., negation and speculation, "*H2A may not be methylated*").
- **Event mention (or event extent)**: text span (e.g., an entire sentence) describing an event and all its components.

An event consists of a trigger and zero or more arguments. Since representing an event as a single trigger word or phrase is explicitly unreasonable [28], most works assume the presence of at least one argument. An event is said to be nested if it has other events in its arguments, flat otherwise. Therefore, it is possible to define the nesting level *k* of an event, starting

TABLE 2. Examples of popular event properties and usual default values.

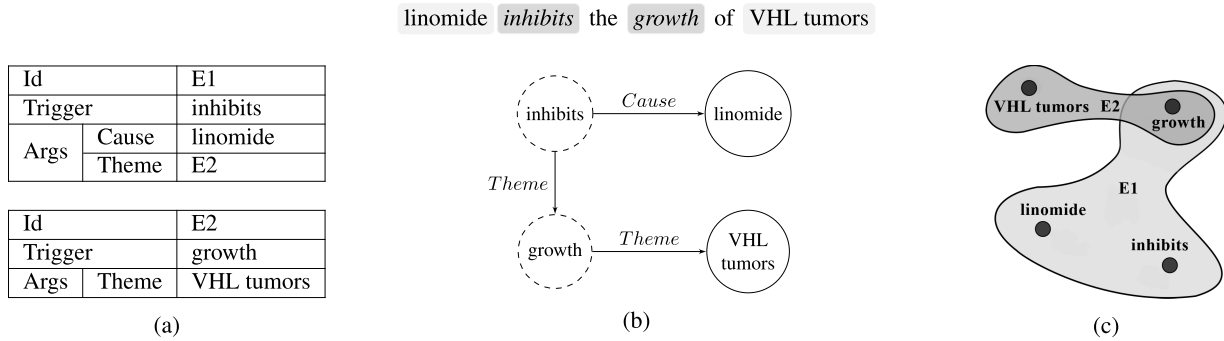| Property	Definition	Possible values
Polarity	The truth value of an event	Positive (*default*) Negative
Tense	The textual anchor time of an event	Past, Future, Present, Unspecified (*default*)
Genericity	Whether an event represents a singular occurrence or not	Specific Generic (*default*)
Modality	Whether an event represents a real occurrence or not	Asserted Other (*default*, e.g., Believed, Hypothetical, Desired)
Speculation	Whether an event is speculated or not	True, False (*default*)
Source	Origin of the knowledge expressed by an event	Current paper (*default*) Other
Manner	The intensity level of an event	High, Low, Neutral (*default*)
Certainty level	The confidence of an event being expressed	L1 (low confidence), L2 (not complete confidence) L3 (high confidence, *default*)
Knowledge type	The overarching information expressed by the event	Investigation, Observation, Analysis, Fact, Method, Other (*default*)
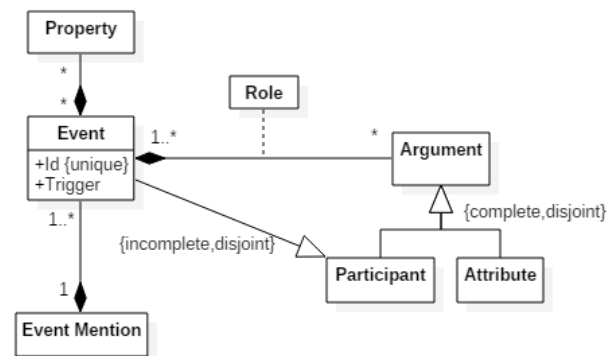
FIGURE 2. ACE-inspired event structure.

from *k* = 1 for flat instances [44]. Nested and recursive event structures are representative of the biomedical domain [44]–[46], where researchers express compound biological processes and molecular events are frequently closely interconnected [40]. For example, in the sentence "*Disruption of curR caused loss of copA expression*", the decreased expression of copA gene is an event that serves as an argument for the parent curR negative regulation event.

The information required by an event structure is encoded with a particular *formalism* (Figure 3). Below, we list the most common ones.

- *Tabular form*. An event table comprises several entries (called event records or slots), one for the trigger and one for each argument as role-value pairs.
- *Direct Acyclic Graph (DAG)*. The nodes are given by the triggers (roots) and the arguments, while the directed edges indicate the argument roles. DAG representations have recently been encouraged by SOTA contributions in biomedical event extraction [44], [47].

linomide *inhibits* the *growth* of VHL tumors

Id	E1	
Trigger	inhibits	
Args	Cause	linomide
	Theme	E2

Id	E2	
Trigger	growth	
Args	Theme	VHL tumors

(a)

(b)

(c)

FIGURE 3. Graphical illustration of three different formalism for a nested negative regulation event: (a) tabular form; (b) DAG; (c) hyperedges.

- *Hyperedge.* Unlike traditional graph edges, hyperedges connect a general set of objects (i.e., vertices) forming a consistent and complete semantic unit. Each event E_i can be modeled as a hyperedge H_i by viewing the interaction among all participating objects as a whole [48]. Since these participants might be of different types, H_i is more general than in classical hypergraph theory [49].
- *Compound value type (CVT) or Mediator.* Adopted by Freebase [50], it is a structured format where each entry consists of multiple fields. CVTs are used to represent n-ary relations where two or more properties can only be interpreted with regard to one another.[5]
- *5W1H tuple.* Moving away from the presented structure, an event can be modeled as a tuple of six elements denoting a 5W1H instance [38].

C. SCHEMA
The events we are interested in (target events or task events) can belong to various predefined types and subtypes. Each of them has its own set of potential participant roles, where the specific classes of participants hang on the event type. For example, an *Infection* event is expected to be described by an *Anatomical* entity with a *Theme* role and by an *Organism* entity with a *Participant* role.

The trigger determines the event type, which in turn delineates the semantic roles of its arguments. As a result, a trigger is usually regarded as the attorney for the corresponding event [28]. Linguists argue that verb↔arguments relationships reflect the event↔participants ones [30]. The idea of relating verbs to specific events or states can also be found throughout the history of philosophy [51].

An event schema (or scenario template) defines the possible arguments for a target event and any entity-type constraints; Table 3 shows an example. An event schema can also define the cardinality of arguments, whose instances may sometimes be optional (secondary) or multiple. Biomedical events typically take a mandatory *Theme*, reflecting a specificity information. The types of events, arguments,

semantic roles, and properties involved in the event schemas are drawn from an externally specified ontology. Long-established ontologies for describing biomolecular events are GENIA Event Ontology [36] and Gene Ontology [53].

III. EVENT EXTRACTION
Event extraction aims at detecting event instances reported in raw text and, if existing, discovering event-related information. It is a long-studied [54]–[56] IE task (rooting back to the 1980s), which has attracted extensive attention among researchers. EE is considered one of the most important research threads in NLP and NLU. Indeed, it requires semantic parsing, i.e., interpreting and distilling free-text event mentions into structured and machine-readable meaning representations (as previously outlined in Section II-B). In other words, EE can systematically organize event knowledge as controlled representations [57]. We note that EE is not the only event-based NLP task, but it is undoubtedly the most popular and well-described.[6]

Biomedical EE (BEE) aims to extract events capturing an interplay between biomedical entities [8], [58], [59]. Extracting and harnessing them is beneficial for medical research and disease prevention [60].

A. COMPARISON WITH RELATION EXTRACTION
Event extraction shares many similarities with RE, ranging from challenges to methodologies. Both IE tasks aim to transform a raw text chunk into structured relational knowledge, not necessarily tied to the underlying syntax. Anyway, sharpening the contrast between them is essential.

In the NLP field, a relation usually refers to a connection between entities in the text. There are many types of relations, namely semantic relations, grammatical relations, and coreferences. The RE goal is to discover semantic ties among several entities. In the RE task, a relation R is generally defined

---

[5]Freebase statistics released on Apr. 23, 2015, stated the presence of approximately 1885 CVT types and 14 million instances. By equating CVTs and events, lots of data are available.

[6]The total number of publications whose title, abstract, and/or keywords refer to event extraction, event clustering, event relation extraction, event embedding, event coreference, and event sequencing are 2,116, 690, 502, 411, 153, 122, and 57, respectively. Data from Scopus® (June 2, 2021), performing queries with the format (*task_name* AND (text OR natural language processing)).

TABLE 3. Example of event schema from Cancer Genetics [52]. The gray labels are shown for organization purposes and are not included in the target types. The format *Arg(T)* indicates an event taking an argument *Arg* which should identify an entity of type *T*. The affixes ?, ∗ and + denote the cardinalities 0..1, 0..N, and 1..N, respectively.

Event Type	Arguments
Pathological	
Mutation	*Theme*(GGP), *AtLoc?*(Anatomical/Pathological), *Site?*
Metastasis	*Theme?*(Anatomical/Pathological), *ToLoc?*(Anatomical/Pathological)
Infection	*Theme?*(Anatomical/Pathological), *Participant?*(Organism)
Molecular	
Gene expression	*Theme+*(GGP)
Transcription	*Theme*(GGP)
General	
Binding	*Theme+*(Molecule), *Site?*(Protein or DNA domain/region)
Regulation	*Theme*(Any), *Cause?*(Any)
Positive Regulation	*Theme*(Any), *Cause?*(Any)
Negative Regulation	*Theme*(Any), *Cause?*(Any)
Planned process	*Theme∗*(Any), *Instrument∗*(Any)

in the form of a tuple $R = r(a_1, a_2, \ldots, a_n)$, where a_i is a named entity or another relation (i.e., nested structure) and r is a predefined relation type [40]. When the arguments are in plain form, R is called unary relation if $n = 1$ (e.g., "phosphorylation of STAT6" → `phosphorylation(STAT6)`), or binary relation if $n = 2$ (e.g., "PAHX interacts with FKBP52" → `interact(PAHX, FKBP52)`). In all other cases, R is called *complex* relation, including n-ary relations.

A notable deficiency of most RE systems is that they focus primarily on directed or undirected[7] binary relations, which results in the extraction of a list of triples connecting only entity pairs. RE can be seen as a classification task that distinguishes whether the candidate entity pair has a semantic relation, possibly identifying its type (if multi-class) [13]. The type becomes necessary even for plain links between drugs and symptoms, where a relationship "cure" rather than "cause" completely alters the semantic meaning. Binary ties may be sufficient for simple IE goals like PPIs, where what is biologically relevant is attesting the existence of some interaction between two proteins, more so than the nature of that interaction or other effects. However, this is not always the case: complex relations are prevalent, and this is especially true in the realm of biomedical phenomena, from molecular-level reactions to cellular-, tissue- and organ-level effects, up to organism-level outcomes. BioNLP-STs and abundant BEE literature have sprung up precisely for filling this gap, mining fine-grained and complex relations to automatically identify and track discoveries and theories expressed in biomedical papers. From a formalization viewpoint, the plain $r(a_1, a_2)$ triplet structure is not enough to fully express text semantics [61], [62], despite being an indispensable representation for the automatic construction of knowledge graphs (e.g., RDF triples) [63], [64]. In biomedical science, it is known that flat triples are often not adequate to represent the complete biological meaning of the original relations [65].

Worse, incorrectly handling n-ary facts leads to extraction errors, such as incomplete, uninformative, or erroneous facts.

EE endeavors to manage higher-order relations involving the equal participation of many different resources. According to the ACE definition, *"whereas a relation expresses a purely binary relationship, an event allows for a trigger and an arbitrary number of arguments (potentially events themselves) having a specific role"*. Therefore, events can be seen as an alternative to the more simplistic and traditional pairwise relations, even beyond complex ones. Table 4 underlines the difference in expressive power between relations and events. Nested and overlapping structures, which occur widely in text, are important to capture inter-event relations such as causality, e.g., a "production" event consequential to a "discovery" event, which in turn results from an "exploration" event.

From a holistic perspective, events are not only a new objective of IE tasks made possible by AI progress and dataset availability, but they represent a different, non-negligible, and under-explored type of knowledge [27], [28], [66], compared to entities and relations.

B. CHALLENGES
Automated EE is still a challenging task for machines. Besides the more elaborated event structure, EE is considerably more arduous than binary RE, and the complexity of biomedical interactions gives further obstructions. Here we list the main barriers observed in our literature review.

- *Directed links*. Under the context of a DAG topology, a prediction is a directed edge with a specific type instead of a plain binary relation [57].
- *N-ary relations*. EE requires capturing n-ary and nested relations rooted in a trigger, defining multi-relational structures. For this purpose, it is possible to factorize n-ary relations into sets of $n - 1$ binary ones. Therefore, pairwise relations can be seen as event building blocks, functional when it is impracticable to extract fine-grained interactions based on information in a text,

---

[7]Non-directionality is appropriate if the roles are the same.

TABLE 4. Difference between binary relation extraction (RE) and event extraction (EE) on two example biomedical sentences. In sentence (a), a traditional RE system ignores the significant contextual information "if over-expressed", which denotes the condition for which the fact `+Reg(Bmi-1,tumorigenesis)` is true. As a result, and by itself, the extracted fact is incorrect. In sentence (b), RE derives two binary incomplete facts instead of a single ternary fact. In both cases, instead, EE correctly captures the expressed interactions.


allowing to derive them later. However, directly employing pairwise RE methods implies low performance (suboptimal solutions) due to the greedy nature of the approach [40], [67]. Let p denote the precision of a binary RE system, the precision deriving from its application to n-ary RE will be p^{n-1}. Moreover, modeling sub-event links in isolation cannot effectively capture relations whose arguments never co-occur in a paragraph. Furthermore, it should also be considered that triggers do not correspond to entities and are likely to be not capturable among the terminal nodes of traditional RE systems (i.e., *entity-entity* links instead of *trigger-entity* or *trigger-trigger* ones). Apart from this, as shown in Table 4, the extraction of binary representations for a complex concept unit may cause a loss of information difficult to recover with a post-processing phase. Thus, the composition of sub-relations to build events constitutes an additional source of complexity and error. Combined with the other peculiarities described in this section, EE generally demands ad-hoc solutions distinguished by those designed for RE tasks.

- *Multiple fine-grained types*. EE may concern many event types, whose arguments may include an extensive range of entity types and semantic roles.
- *Links among arguments*. EE may require the identification of links among the arguments of a same event. For example, a *Binding* event type can take the following argument role types: *Theme*, *Theme2*, *Site*, and *Site2*, where the number "2" is attached to differentiate argument linkages (i.e., *Theme↔Site*, *Theme2↔Site2*) [68].
- *Deeper understanding*. Events have different components, while natural languages have semantic ambiguities, a variety of narrative styles, and a deluge of syntactic constructions. Multiple events may exist in the same sentence (≈27% in ACE 2005) [69], [70]. The same extent or trigger might express different events depending on the sentential context. Similarly, an event might be described by completely different expressions

and trigger words. Even within the same document, an event might be referenced by various extensions (*event coreference*) [71]. Triggers can overlap by sharing a text span (*event overlapping* or *double tagging problem*), and the same goes for entities [44]. Within the same event type, argument roles are nonidentical and often not interchangeable: in a tuple or predicate view, event argument order matters. The same argument can also take part in multiple events with different roles (*role overlapping* [72]). Arguments may be indirectly mentioned (e.g., "both proteins"), demanding entity mention coreference resolution. Failures of coreference resolution are a common and oft-cited cause of false negatives in IE from the scientific literature.

- *Long-range dependencies*. An event may involve arguments that are distant from each other or from the trigger, peculiarly frequent in long documents such as biomedical publications. Compared to RE, an event extent might be composed of many sentences, and it is hard to find all arguments of one event instance in a single sentence (arguments-scattering) [73]. Simply put, events do not necessarily constitute continuous spans of text, and their components are fragmented.
- *High-level linguistic phenomena*. Linguistic modality is frequently used in natural language to express uncertainty with respect to events. Many NLP applications seek to distinguish positive/factual statements from negative/speculative ones [74], [75]. The interpretation may differ even under the same trigger and arguments. For example, *"It is known that"* (generally accepted fact) / *"We examined whether"* (under examination) / *"The results suggest that"* (speculation) *"the narL gene product activates..."* [42]. Such information is crucial in the medical domain, where it can facilitate more accurate IE and search [76], [77]. For example, suppose we pose a question to a downstream question answering model, such as *"Did drug X cause side effect Y?"*. In that case, assuming trustworthy source text, the answer should
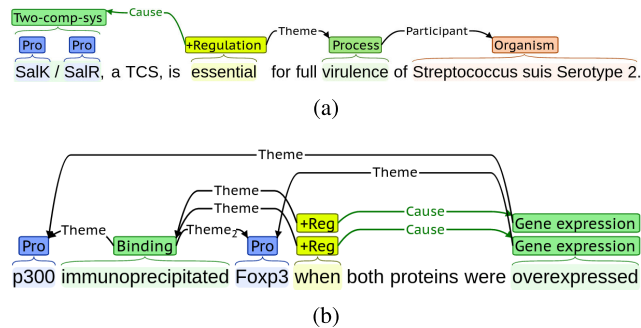
FIGURE 4. Advanced events: (a) entity overlapping (simplified from PMC2358977); (b) entity coreference resolution and double tagging (PMC1447668). Event DAG visualizations created with `BRAT` [81].

be different depending on the evidence that the system has observed: *"Drug X caused side effect Y"* [yes] or *"Drug X is unlikely to have caused side effect Y"* [uncertain]. It follows that event modifiers are as difficult to extract as they are essential, especially on an application side.

Biology amplifies these requirements, outlining a more complicated application territory for EE. To raise the difficulty are the high number of types (entities, roles, and events) involved, an above-average number of interacting participants, the high occurrence of nested and overlapping events, the greater length of source documents compared to that of other domains, and the regular presence of meta-knowledge. Furthermore, different from the general news domain, biomedical IE typically necessitates much broader domain-specific knowledge [78]. Biomedical documents, either formal (e.g., scientific papers) or informal (e.g., clinical notes), are written for domain experts. As such, they contain many specialized terms, abbreviations, and acronyms. Many shortcomings also arise in biomedical coreference resolution, with longer chains and referring entities/events hardly to spot without a domain-specific comprehension of semantic classes [79], [80].

Figure 4 shows some examples.

C. CORPORA AND EVALUATION DATASETS

A series of corpora and evaluation datasets have been proposed to inspire EE study, providing a direct point of comparison between different approaches and systems. Generally, standardized corpora construction is supported by public evaluation programs, where labeling is manually performed by experts, and events are treated as primary text-bounded annotation objects. Gold-standard event annotations can be used as benchmarks or training datasets. Each dataset refers to predefined target events on one or more topics, in line with a set of fixed schemas. They can fully or partially exploit the event structure introduced in Section II-B, e.g., no nested or overlapping definitions. Historically, two domains that have contributed most to the development of NLP for EE are news and biology.

1) NEWSWIRE DOMAIN

The Message Understanding Conference, held seven times from 1987 to 1997 (MUC1-7) [82], has been recognized as the first public evaluation program for IE systems on pre-established tasks. Automatic Content Extraction (ACE) [83] is the most influential EE program up to now [26]. Proposed by NIST in 1999 to develop technologies for extracting and characterizing textual meanings, ACE supplies ground truth labels and evaluation tools for different kinds of extraction targets (or "knowledge elements"): entities, values, relations, and events (added in 2005). Anyhow, the ACE 2005 (English) event corpus [84] is as widespread as limited: flat event structures, >95% single-word triggers [85], 57% ambiguous trigger words [86], ≈60% event types with less than 100 labeled samples [87], coarse event schemas [88], non-overlapping events, sentence-level extraction, and highly skewed type distributions [89]. ERE standard [90] is a lighter-weight version of ACE with similar event schema definitions, pursuing the goal of making annotation easier and more consistent across annotators. TAC-KBP [91] is another newswire dataset commonly used in current EE research; it is more oriented to KB population and considers fewer types for more efficient dataset creation/evaluation, also handling event overlapping.

2) BIOMEDICAL DOMAIN

After centering on the general language domain, the research spotlight has recently widened to the biomedical field, motivated by the ongoing series of BioNLP shared tasks (STs). The BioNLP-ST series represents the main community effort to address BEE. BioNLP-ST'09 [92] fueled the interest in this field, introducing a dataset as a tractable subset of the GENIA event corpus [68], made up of 1,210 MEDLINE abstracts and 13,588 biomolecular events. BioNLP-ST'11 [93] was organized under the "Generalization" theme: (i) the GENIA dataset was extended with 14 PMC full-text articles; (ii) coreference, entity relations, and gene renaming were included as BEE-supporting tasks; (iii) new BEE corpora on uncovered topics were inserted. In its third edition, BioNLP-ST'13 [94] simulated a more realistic scenario, pursuing the grand theme of "KB construction" with a semantic web vision: (i) the previous GENIA corpus was replaced with a new version using only up-to-date full papers (making comparison difficult); (ii) new BEE tasks were introduced, answering the biologists' needs and testing scalability to richer ontologies. Afterward, BioNLP-ST'16 [95] still moved in this direction, clarifying the impact of BEE tasks on real-world contexts and not relying solely on intrinsic evaluation (one of the major criticisms of that time [96]). With the latest 2019 edition, the series made a significant change, opening task registration to reduce the bias caused by a small number of organizers, thus offering anyone the opportunity to be a task provider. Since then, the title has changed to BioNLP *open* shared tasks (BioNLP-OSTs) [97].

To minimize the time investment necessary to participate in the shared tasks, some BioNLP competitions provide

participants with annotations from SOTA automated tools and linguistic resources, like BioLemmatizer [98].

In all, GENIA is an anchor point for the BioNLP-ST series. GENIA'11 events, for example, have properties not embraced by the ACE 2005 specification: 37.2% nested, 6.0% inter-sentence, 36% 10+ word separation between triggers and arguments [99], 24% overlapping entities [100], pure entity-based arguments, etc.

Other than BioNLP-ST competitions, many biomedical corpora have been compiled by different research groups; each varies in terms of domain, size, and annotation schema. A closer breakdown of the leading BEE resources and their descriptive statistics can be found in Table 5. Supplementary corpora collections are available on the Web[8],[9] Among these, we mention the BioCreative BEL dataset [101]–[103] for its distinctive features, where composite sentence-level causal and correlative relationships are encoded with a formal language using functional predicates. For example, the BEL statement `p(HGNC:TIMP2) -| cat(p(HGNC:MMP2))` indicates that the human protein TIMP2 decreases the catalytic activity of MMP2. Although the BEL task is similar to those of BioNLP, it focuses not only on the extraction of interactions between biomedical entities but also on NER and entity normalization based on namespaces derived from public repositories. So, BEL data do not include gold entity offsets, types, and identifiers, often leading systems to integrate multiple NLP tools.

3) DATASET LIMITATIONS

Compared to other IE tasks such as name tagging, annotations for events are more costly because of their complex structure and the required rich label space. As the event annotation process is cost-prohibitive (i.e., time-consuming, labor-intensive, domain knowledge demanding, and based on task-specific guidelines), resources are few, small, and have low coverage. For instance, annotating the 9,372 sentences of the original Genia Event Corpus took 1.5 years with five part-time annotators and two coordinators [36]. The disadvantages around hand-labeled data unveil the lack of comprehensive event-related KBs [26]. Crucially, typical training data for EE include fewer than 300 annotated abstracts (Table 5). Another critical point concerns the common presence of unbalanced training data, especially for larger datasets, which can greatly degrade recognition performance. The maldistribution of samples is quite serious in EE benchmarking, where a large portion of event types contain relatively few training instances. For example, among the 19 event categories and 6,667 event annotations provided by MLEE, the most annotated class has nearly 1000 instances, while the less annotated one has only 10. Moreover, there are eight categories whose annotations are less than 100 [113]. Hence, in addition to the task difficulty and the modeling challenge,

insufficient training data and imbalanced event types hinder the learning process and represent significant obstacles for democratizing EE.

4) OVERLAPPING ANNOTATIONS

Along with the issues delighted in Section III-C3, many entity and event types are annotated in multiple corpora, blurring their ontological border. As for the biomedical domain, many systems [43], [114] build on the assumption that types sharing the same label (e.g., "Binding") in two corpora denote the same semantic type and are annotated in a broadly compatible manner. Regardless of the domain, this assumption is expected to hold for several resources due to the shared ontological foundations granted by an ontology-based schema definition. Thus, datasets with overlapping annotation scopes drawing their event types from the same ontology are often deemed comparable. This is typical of the event field and very different from other NLP tasks and datasets which do not share ontologies, types, or annotation criteria. Annotation differences can also occur when an event type is identically annotated in two different corpora. For instance, while a *Localization* event involves only proteins in GE11, the same event type on MLEE also admits genes and cells with a *Theme* role.

5) DATASET FORMAT AND EVALUATION METRICS

Whilst all hinges on the examined task, EE assessment ordinarily rests on instance-oriented standard precision, recall, and F1-score metrics.

In ACE 2005, the evaluation is done at the level of single arguments and their roles, without considering completeness. In contrast, GENIA tackles just the correctness of complete events, splitting the nested ones.

The standoff annotation format adopted by BioNLP-STs is called *.a**, where labels are connected to spans of text through (*start*, *end*) character offset pairs. To focus only on targeting EE, text documents are complemented by input annotation files with gold entities (.a1), while the expected extraction output is reported in target annotation files (.a2). Event equality holds between any two events when (i) the event types are the same, (ii) the event triggers are the same, and (iii) the arguments are fully matched. Accordingly, an event trigger or argument is correctly identified if its offsets match those of a gold-standard annotation, and it is correctly classified if its type or role also fits the ground truth. The required match could be strict or approximate, allowing the predicted spans to differ from the gold ones by single words (i.e., a given span is accepted if it is entirely contained within an extension of the gold span by one word both to the left and to the right). Normally, for a nested event to be valid, the events it refers to should also be strictly correct. The "approximate recursive matching" mode relaxes this requirement so that an event can match even if the sub-events have at least valid Theme arguments (partial matching). It is worth noting that annotations for the test sets of the BioNLP-ST corpora are not provided; instead, users who want to perform evaluation need

---

[8]BioCreAtIvE: Critical Assessment of Information Extraction in Biology. http://www.biocreative.org/ (accessed Nov. 23, 2021).

[9]NaCTeM: The National Centre for Text Mining. http://www.nactem.ac.uk/ (accessed Nov. 23, 2021).

TABLE 5. Summary of the main newswire and biomedical event extraction corpora, in intra-class publication order. The number of documents is detailed in [train, validation, test] or [train, test]. % nested events refers to the test set.

Corpus	Domain(s)	#Documents	Annotation Schema	#Annotations
NIST ACE 2005 English Event Corpus [84]	Spoken and written media sources on broadcast news, telephone conversation, newswire, usenet, and weblogs	599 documents crawled between 2003 and 2005	8 event types, 33 event subtypes, 35 argument roles, 4 modifiers ↪ polarity, tense, genericity, modality	≈6,000 events, 5,272 triggers, 9,612 arguments
NIST TAC-KBP 2016 [91]	Newswire and discussion forums	≈90K documents (Chinese, 30K-English, and Spanish) ↪ 505 annotated	8 event types, 18 event subtypes 3 modifiers ↪ belief-type, sentiment sarcasm	≈4,000 events
Bio Information Extraction Resource (BioInfer) [104]	Interacting proteins	2,240 sentences (from abstracts)	110 entity types, 60 event types, 1 modifier ↪ negation	6,349 entities, 2,662 events
Genia Event Corpus [68]	Humans blood cells transcription factors	1,000 abstracts	35 terminal entity types, 35 event types	93,293 entities, 36,114 events
Gene Regulation Event Corpus (GREC) [105]	Gene regulation and expression within E. coli and human species	240 abstracts	174 entity types, 14 role types, 353 trigger verbs, 19 event types	5,026 entities, 3,067 events
BioNLP-ST'11 Genia Event 2011 (GE2) [106]	Humans blood cells transcription factors	1,210 abstracts, 14 full papers [[800,150,260], [5,5,4]]	2 entity types, 9 event types, 2 modifiers ↪ negation, speculation	21,616 proteins, 18,047 events, 45.17% scattered args, 53.83% nested events
Epigenetics and Post-translational Modifications (EPI) [107]	Epigenetic change and common protein post-translational modifications	1,200 abstracts [600,200,400]	2 entity types, 14 event types, 2 modifiers ↪ negation, speculation	15,190 proteins, 3,714 events, 369 modifications, 8.27% nested events
Infectious Diseases (ID) [45]	Two-component regulatory systems	30 full papers [15,5,10]	5 entity types, 10 event types, 2 modifiers ↪ negation, speculation	12,740 entities, 4,150 events, 214 modifications, 41.50% nested events
Multi-Level Event Extraction (MLEE) [108]	Blood vessel development from the subcellular to the whole organism	262 abstracts [131,44,87]	16 entity types, 19 event types	8,291 entities, 6,667 events, 42.04% scattered args, 53.17% nested events
GENIA-MK [42]	Same as Genia Event Corpus (meta-knowledge assignment)	1,000 abstracts [800,200]	5 modifiers (+ 2 inferable) ↪ knowledge type, certainty level, polarity, manner, source	36,858 events
BioNLP-ST'13 Genia Event 2013 (GE3) [109]	Humans blood cells transcription factors	34 full papers [10,10,14]	2 entity types, 13 event types, 2 modifiers ↪ negation, speculation	12,068 proteins, 762 entities, 9,364 events, 58.89% nested events
Cancer Genetics (CG) [52]	Cancer biology	600 abstracts [300,100,200]	18 entity types, 40 event types, 2 modifiers ↪ negation, speculation	21,683 entities, 17,248 events, 1,326 modifications, 50.76% nested events
Pathway Curation (PC) [110]	Reactions, pathways, and curation	525 abstracts [260,90,175]	4 entity types, 23 event types, 2 modifiers ↪ negation, speculation	15,901 entities, 12,125 events, 571 modifications, 57.57% nested events
Gene Regulation Ontology (GRO) [111]	Human gene regulation and transcription	300 abstracts [150,50,100]	174 entity types, 126 event types	11,819 entities, 5,241 events
Gene Regulation Network (GRN) [112]	Sporulation in Bacillus subtilis	201 sentences (from abstracts)	12 entity types, 9 event types, 6 interaction* types *entity-event relations	917 entities, 495 events, 334 interactions

to upload their predictions to the task organizers' servers. A closer look to .*a** is available in Figure 6.

Unlike the modeling and labeling aspects, we notice that EE evaluation has attracted very little research attention. Outside of BioNLP-STs, many works still employ traditional metrics to calculate an average score based on the predictions generated for each expected role (i.e., slot-level event table comparison) [115]. Despite being indicative of model performance to some extent, averaged metrics can be pretty misleading to downstream applications managing event instances as a whole, where one wrongly identified argument can substantially alter the whole meaning of an event instance. This event-as-a-whole demand is common across many domains, such as health, legislation, and finance. To mitigate this gap, Zheng *et al.* [116] proposed two new evaluation metrics to explicitly penalize wrongly identified event arguments, namely NOFP and NOFPFN. If we only care about the correctness of predicted events, NOFP treats any entity-level false-positive (FP) error as an event-level FP error. If we further pursue event completeness, NOFPFN permits neither FP errors nor false-negative (FN) errors as regard entities.

D. LITERATURE REVIEW

EE research evolves from pattern matching to deep neural models, from sentence-level contexts to entire documents. In this section, we point out and review the different approaches proposed over time, which are illustrated in Figure 5 in the form of hierarchical bibliographic categories. For the sake of clarity, we distinguish manually designed rule-based systems (which rely on explicit and static models) from learning systems (which create their own models). While Figure 5 is designed to characterize the EE space given the most valuable dimensions from a practical side, Table 6 makes easier for the reader to identify the main information of each BEE contribution. We intend to convey a broad-spectrum view on BEE, linking research in the biomedical field with that conducted in general EE. During the discussion, solutions applied to the biomedical domain are explicitly indicated; if not specified, we refer to an evaluation on a different domain (e.g., news or general).

1) CLOSED-DOMAIN AND OPEN-DOMAIN

Closed-domain EE (henceforth referred to as "CDEE") uses predefined event schemas to discover and extract desired events from the text. In a nutshell, its goal is to fill the expected slots (table formalism) or expand the current paths along pre-established directions (graph formalism). From an ontological point of view, CDEE task can be oriented to identify only terminal event types or even intermediate ones.[10] Coupled with ACE 2005, BioNLP-STs have nourished the community to develop a large number of CDEE systems, resulting in a landscape full of contributions. CDEE heavily relies on training data annotation schemas, but these templates are inherently rigid and inflexible [27]. As the name suggests, CDEE solutions are generally tailored to specific domains, are difficult to deflect from their original purpose, and their effectiveness is limited to the semantic types they are trained to identify. Such trained models hardly extract large-scale events [25], and their performance may drop significantly when the schema changes, even within the same domain [247]. Yet, most situations admit an extensive range of variation and novelty, sometimes blending multiple events. Domain adaptation approaches usually filter out redundant and contradicting labeling across different corpora, subsequently merging them to train a single combined model [24]. Real-world applications tend to be open and evolve promptly; thus, there can be numerous new unseen event types. Naturally, there cannot be a schema for everything: an exhaustive enumeration treating all possible event types and scenarios would be impractical[11] [28].

Open-domain EE (henceforth referred to as "ODEE") does not assume predefined event schemas. It aims to detect all kinds of events from texts and, in most cases, also cluster similar events via extracted event keywords (i.e., characterizing words/phrases) [25]. Unlike CDEE, event schemas are not received as input, but their learning can be a supplementary output, mitigating the adaptability pitfall. ODEE systems rely mainly on unsupervised extraction strategies to automatically discover possible relations of interest at scale. On the other hand, the openness for the domain and the types of events to be extracted (together with their constituent elements) prevents the definition of an established annotated corpus as a gold standard dataset for objective and reproducible cross-system comparisons. As a consequence, to date, ODEE systems were predominantly evaluated by hand or with extrinsic approaches, thereby ignoring one of the principal aims of ODEE: scalability to large amounts of heterogeneous text.

ODEE is related to the open-domain information extraction (Open IE or OIE) paradigm, where the goal is extracting all types of relation tuples (e.g., $\langle arg1, rel, arg2 \rangle$) found in a text, without being limited to a small set of target relations known in advance. While most OIE systems are based on binary relations, some recent solutions can extract tuples of arbitrary arity, sometimes also with modifiers and attributes [250], making it possible to weakly compare with ODEE solutions. Since OIE systems are not aligned with an ontology, arguments and semantic relations are restricted to untyped phrases. For example, given the sentence "proton pump inhibitors may cause abdominal or chest pain", a traditional OIE system extracts a tuple like $\langle$*proton pump inhibitors*, *may cause*, *abdominal or chest pain*$\rangle$.

2) SUBTASKS AND EXECUTION MANNER

Ordinarily, CDEE can be broken down into four subtasks.

---

[10]Coarse-grained types are typically used as targets for rare events without appropriate fine-grained types.

[11]Interestingly, in some general-domain knowledge bases (like *ATOMIC* [248] and $ATOMIC^{20}_{20}$ [249] for commonsense reasoning), an event is directly represented as a sentence instead of a structured format.
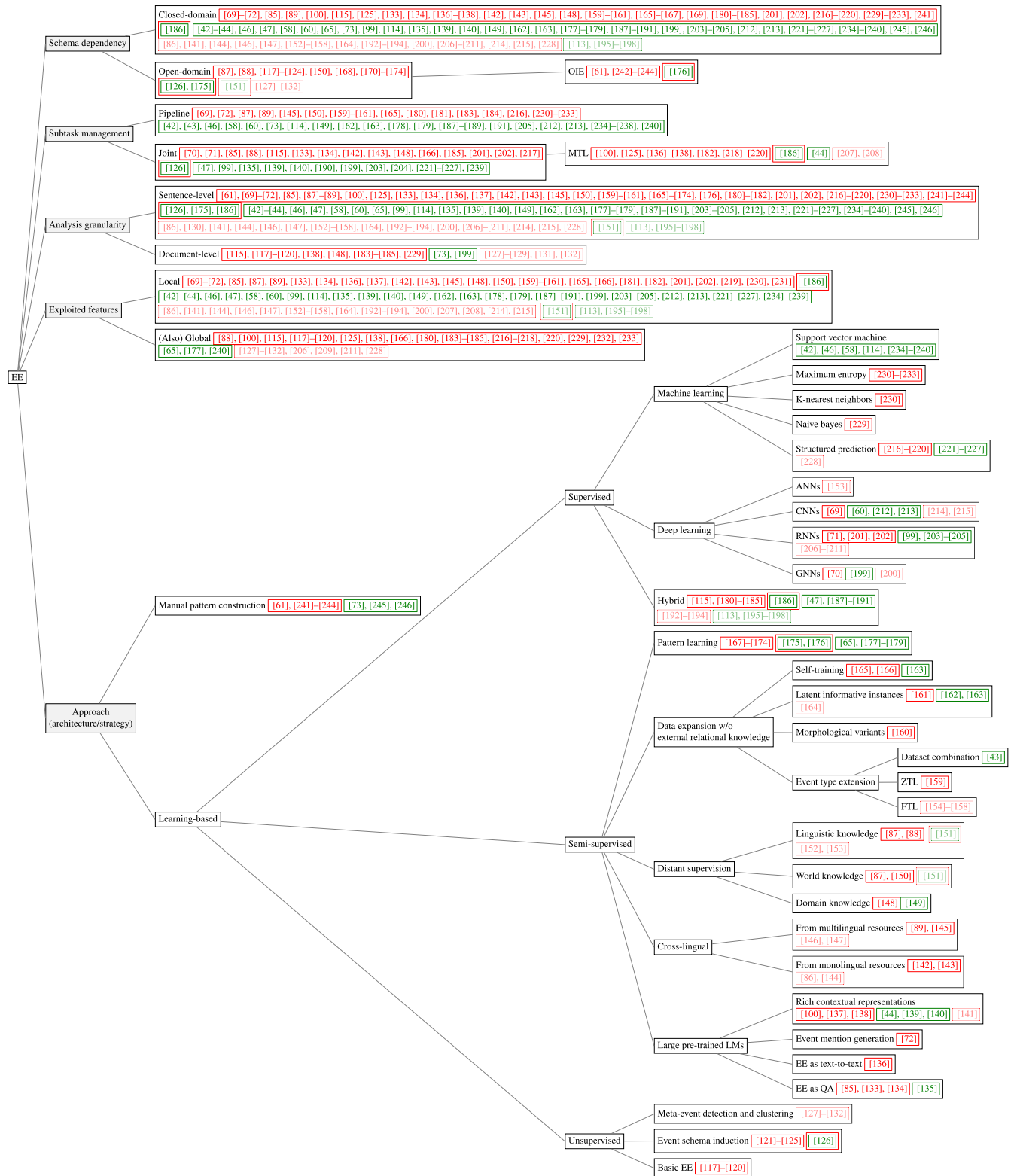
FIGURE 5. Taxonomy of the reviewed literature and main approaches identified for event extraction. The gray nodes represent different and partially overlapping classification criteria. In agreement with task specificity or evaluation domain, references boxed in green and red correspond to biomedical and non-biomedical solutions, respectively. Lighter references stop to event detection. Supervised works are distinguished by their architecture, while semi-supervised and unsupervised ones by their key strategy. Subtask management does not apply to event detection and rule-based systems, while features refer only to solutions requiring learning.

TABLE 6. Summary of the reviewed biomedical event extraction systems, their characteristics, and evolution over time.

	Schema depend.		Subtask management			Analysis granul.		Exploited features		Approach				Description	Year
	Closed-domain	Open-domain	Pipeline	Joint	MTL	Sentence-level	Document-level	Local	(Also) Global	Manual Pattern C.	Supervised	Semi-supervised	Unsupervised		
Patwardhan and Riloff [175]		✓				✓		✓				✓		Pattern learning + semantic affinity	2007
Cohen et al. [245]	✓					✓				✓				Human-crafted patterns (ontology-driven)	2009
Kilicoglu and Bergler [179]	✓		✓			✓		✓				✓		Pattern learning w/ dependency paths	2009
Björne et al. [234]	✓		✓			✓		✓			✓			SVM	2009
Riedel et al. [221]	✓			✓		✓		✓			✓			Structured prediction	2009
Miwa et al. [236]	✓		✓			✓		✓			✓			SVM	2010
Casillas et al. [246]	✓					✓				✓				Morpho-syntactic and semantic patterns	2011
Minh et al. [178]	✓		✓			✓		✓				✓		Pattern learning (POS tags near proteins)	2011
Björne and Salakoski [58]	✓		✓			✓		✓			✓			SVM	2011
Riedel et al. [222]	✓			✓		✓		✓			✓			Structured prediction	2011
McClosky et al. [223]	✓			✓		✓		✓			✓			Structured prediction	2011
Riedel et al. [225]	✓			✓		✓		✓			✓			Stacked structured predictions	2011
Bui and Sloot [65]	✓					✓			✓		✓			Pattern learning w/ structured mapping	2012
Miwa et al. [42]	✓		✓			✓		✓			✓			SVM	2012
Miwa et al. [114]	✓		✓			✓		✓			✓			SVM + coreference resolution	2012
Vlachos and Craven [224]	✓			✓		✓		✓			✓			Structured prediction	2012
McClosky et al. [226]	✓			✓		✓		✓			✓			Structured prediction	2012
Miwa and Ananiadou [46]	✓		✓			✓		✓			✓			SVM	2013
Björne and Salakoski [235]	✓		✓			✓		✓			✓			SVM	2013
Hakala et al. [240]	✓		✓			✓			✓		✓			SVM + EVEX-based re-ranking	2013
Miwa et al. [43]	✓		✓			✓		✓				✓		Data expansion (dataset combination)	2013
Liu et al. [239]	✓			✓		✓		✓			✓			SVM pairwise extraction and extension	2013
Bui et al. [177]	✓					✓			✓		✓			Pattern learning w/ structured mapping	2013
Pham et al. [187]	✓		✓			✓		✓			✓			Hybrid (Pattern matching + ML)	2013
Liu et al. [188]	✓		✓			✓		✓			✓			Hybrid (Pattern matching + ML)	2013
Venugopal et al. [190]	✓			✓		✓		✓			✓			Hybrid (MLN + SVM)	2014
Björne and Salakoski [237]	✓		✓			✓		✓			✓			SVM	2015
Liu et al. [189]	✓		✓			✓		✓			✓			Hybrid (graph-based patterns + ML)	2015
Munkhdalai et al. [163]	✓		✓			✓		✓				✓		Data expansion (self-training + latent informative instances)	2015
Lever and Jones [238]	✓		✓			✓		✓			✓			SVM	2016
Li et al. [203]	✓			✓		✓		✓			✓			Bi-LSTM	2016
Li et al. [227]	✓			✓		✓		✓			✓			Structured prediction	2016
Huang et al. [126]		✓		✓		✓		✓					✓	Event schema induction (symbolic + distributional semantics)	2016
Han et al. [162]	✓		✓			✓		✓				✓		Data expansion (latent informative instances)	2016
Wang et al. [213]	✓		✓			✓		✓			✓			CNN	2017
Rao et al. [149]	✓		✓			✓		✓				✓		Distant supervision (domain knowledge)	2017
Björne and Salakoski [212]	✓		✓			✓		✓			✓			CNN	2018
Li et al. [205]	✓		✓			✓		✓			✓			Bi-GRU	2018
Li et al. [99]	✓			✓		✓		✓			✓			Tree-LSTM	2019
Li et al. [60]	✓		✓			✓		✓			✓			Multi-pooling CNN	2020
Zhu and Zheng [47]	✓			✓		✓		✓			✓			Hybrid (CNN + RNN)	2020
Trieu et al. [44]	✓			✓	✓	✓		✓				✓		SciBERT + custom layers	2020
Chung et al. [73]	✓		✓				✓			✓				Patterns w/ context triggers	2020
Wang et al. [135]	✓			✓		✓		✓				✓		EE as QA	2020
Huang et al. [139]	✓			✓		✓		✓				✓		SciBERT + GNN	2020
Ramponi et al. [140]	✓			✓		✓		✓				✓		EE as sequence labeling	2020
Yu et al. [204]	✓			✓		✓		✓			✓			Bi-LSTM + Tree-LSTM	2020
Zhang et al. [186]	✓				✓	✓		✓			✓			Hybrid (Bi-LSTM + GCN)	2021
Abdulkadhar et al. [191]	✓		✓			✓		✓			✓			Hybrid (EnsembleSVM + graph kernels)	2021
Zhao et al. [199]	✓			✓			✓	✓			✓			GNN (Stacked HANN layers)	2021

1) *Trigger detection (TD).* Detect whether an event exists and, if so, the textual chunk that acts as event lawyer.

2) *Trigger classification (TC).* Classify the trigger as one of the given event types.

3) *Argument detection (AD).* Predict event arguments.

4) *Argument role classification (AC).* Associate the event trigger with its arguments and classify the semantic roles contingent on the recognized event type (i.e., edge detection, reckoning with a graph formalism).

Event detection (ED) generally refers to TD + TC. A post-processing stage is usually present to improve and complete the candidate events, optionally comprising modality identification. TD dominates the whole extraction process since it strongly affects the following steps [108]. Some studies have shown that trigger misclassification causes ≈70% of all errors in ED [251].

The four subtasks can be executed in a *pipeline* manner or a *joint* manner. In the first case, multiple independent classifiers are trained, and the output of one serves as a part of the input for the next (i.e., TD → TC → AD → AC). In the second case, one model performs multi-task classification simultaneously in an end-to-end fashion. The "one for all" philosophy of EE joint architectures can also be extended to additional NLP tasks (e.g., NER, coreference resolution, RE, event relation extraction), allowing each to improve the other and opening doors to *multi-task learning (MTL).*[12] For instance, knowledge of coreference relationships can help infer the type of entity mentions otherwise tricky to classify.

Pipeline architectures — oft-times a consequence of reliance on pre-existing NLP toolkits which interrupt backpropagation — simplify the problem but ignore the inter-dependencies between subtasks and suffer from error propagation (i.e., error inheritance and amplification from upstream classifiers, also known as "snow ball" effect). Joint inference "breaks the pipeline", benefits from parameter sharing, and allows the flow of deeper information between subtasks both forward and backward, mutually improving local predictions and leading to higher precision [207].

3) MANUAL PATTERN CONSTRUCTION

Analogously to other NLP tasks, early attempts on EE adopted *human-crafted* templates to match the events, implementing pipeline-based symbolic solutions [241], [245], [246]. A template can be constituted by patterns or rules encoding domain and linguistic knowledge; constraints may involve specific lexical items or semantic classes. Many pattern-based EE systems have been released for different application domains, including biology (with ontology-driven semantic grammar [245], as well as morpho-syntactic and semantic patterns [246]). For example, the word "expression" following a gene name generally implies an expression event for that gene.

Various ruleset-based OIE systems have been constructed and experimented with for ODEE [61], [242]–[244]. Akbik and Löser [61] proposed KrakeN to identify n-ary facts with hand-written extraction rules over typed dependency parses. Similarly, Mesquita *et al.* [242] presented Exemplar

to detect relation triggers and multiple arguments connected to them, assigning each argument its corresponding SRL role (e.g., subject, direct object, or prepositional object). Buthani *et al.* [243] built NESTIE to capture nested relations (e.g., ⟨⟨*arg*1, *rel*1, *arg*2⟩, *rel*2, *arg*3⟩) with a set of manually defined rules over dependency parse trees, allowing for a more accurate representation of the input sentence meaning. De Vroe *et al.* [244] introduced MONTEE, an n-ary modality-aware ODEE system, allowing the distinction between events that took place, did not take place, and for which there is a degree of uncertainty. The authors inferred relations of higher valences by combining sets of binary relations (e.g., "arg1-predicate-arg2-preposition-arg3" from "arg1-predicate-arg2" and "arg2-preposition-arg3"). They used a modality lexicon and a set of dependency graph patterns in tandem to assign modality categories to event tuples.

While pattern matching can achieve high extraction accuracy if carefully refined by experts, it suffers from noteworthy issues: (i) it is expensive to build and maintain; (ii) it is pretty rigid, brittle, and hardly generalizable across domains; (iii) it struggles to handle semantics and exigent NLU tasks. CDEE evaluation results from BioNLP-ST'09 and BioNLP-ST'11 have shown that rule-based systems generally achieve moderate effectiveness, with high precision, low recall, and better predictions for simple events. Since most computation is for matching pre-established rules against text, such systems exhibit good efficiency.

4) SUPERVISED LEARNING

So far, most EE methods follow the supervised learning paradigm, which assumes sufficient event mentions and ground truth labels within annotated corpora (see Section III-C). Machine learning (ML) and deep learning (DL) techniques largely belong to this group, and their basic idea is almost the same: learning EE classifiers from training data and applying them to new text [26] (Figure 6).

a: FEATURE TYPES AND EXPLOITED INFORMATION

Features need to be appropriate if we want our supervised EE system to be effective.

Many contributions depend on local information at the phrase or sentence level (i.e., *local features*) [190], [209]. Most of them can be obtained via off-the-shelf NLP tools and are generally divided into three types [26], [69]: *lexical* (e.g., word- and subword-level n-grams, lemmas, POS tags), *syntactic* (e.g., dependency parsing relations), and *semantic* (e.g., synonyms, event/entity types). Contextual word embeddings fit in the lexical-semantic class. Additional semantic entity knowledge can provide useful information for ED [28]. Depending on their importance for specific EE subtasks, another frequent distinction is between *trigger* and *argument* features [217].

Local features may be insufficient to resolve ambiguities in identifying some event types, which can be missed or spuriously tagged [28], [211], [218], [228], [233]. A few approaches cross sentence boundaries, eventually going

---

[12]MTL depicts a kind of transfer-learning paradigm, which enhances generalization by using the domain information in the training signals of related tasks as an inductive bias [252]. In this paper, we use "MTL" to indicate models trained both on EE and external tasks.
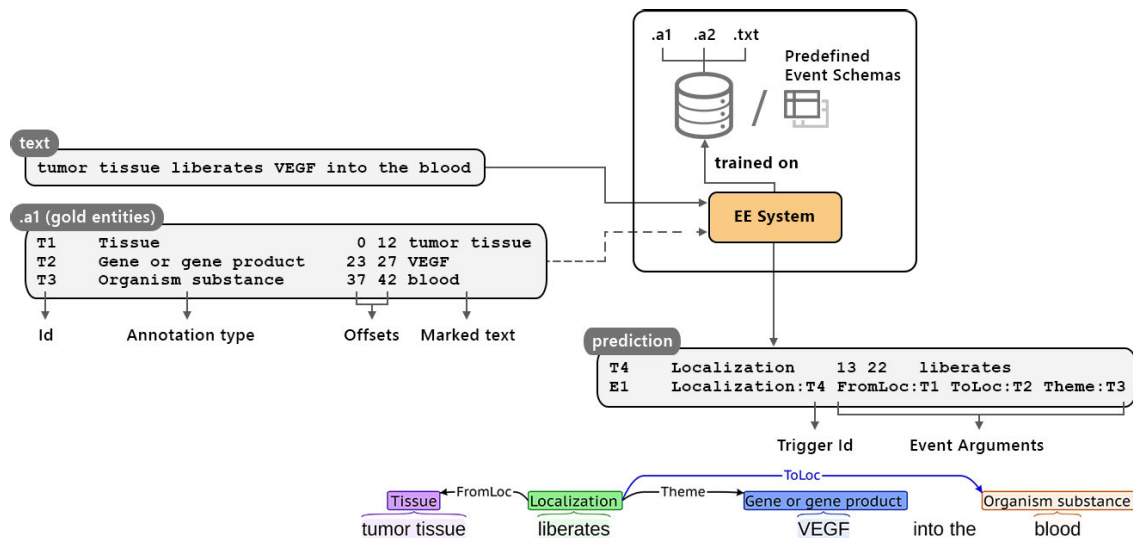
FIGURE 6. Example of a supervised closed-domain biomedical event extraction process with *.a** annotation format. The reliance on gold entities is optional and hinges on the specific system.

beyond documents, to exploit richer contexts (i.e., *global features*), showing improvements even for sentence-level EE. High-level features can be human-designed (e.g., dependencies among triggers and arguments [232], [233], temporal correlations [180], discourse properties [216], event-to-event relations [88]) or not (e.g., document embeddings [206], [209]).

b: MACHINE LEARNING

Traditional ML approaches rely on human-engineered features, calling for linguistic knowledge and domain expertise. A multitude of ML systems for EE adopted pipeline architectures and classifiers grounded on kernel functions, like support vector machines (SVMs). As for the biomedical literature, the Turku Event Extraction System (TEES SVM [58], [234], [235], [253]) and EventMine [46], [236] — proactively evolved in different versions — have represented the epitome of this line of work, with the ability of automatically adapting to various tasks [43], [114], [237], [254]. TEES SVM stood on text-bound graphical structures, reformulating BEE as an event graph construction problem, decomposed in consecutive and straightforward node/edge classifications. TEES SVM has been employed in large-scale TM applications, e.g., extracting events from 19 million unlabeled PubMed abstracts [255], or generating a collection of over 40 million biomolecular events also from PMC full-text articles (called EVEX) [240], [256]. EventMine (further extended to EventMine-MK [42]) is a similar SVM-based system developed by NaCTeM, which lately dethroned TEES SVM on the original 2009 GENIA Corpus, plus ranking first and second in the PC and CG BioNLP-ST'13 tasks. Building upon the success of TEES, VERSE [238] offered complete control over the specific semantic features to use for classification, potentially originating different parameter sets for

various tasks. Hakala *et al.* [240] re-ranked the output of TEES by incorporating global features from the EVEX resource as world knowledge. Alternative procedures resorted to k-nearest neighbor and maximum entropy learners [230], [231].

Some ML contributions made use of global features to endow sentence-level EE systems with more information derived from statistical hypotheses (e.g., trigger and argument role consistency within topically related documents, frequent co-occurrence of certain event types) [216]–[218], [228], [232], [233].

Joint ML classification models allowed a great advancement in terms of overall performance, with efforts geared on structured learning (e.g., perceptrons, Markov logic, conditional random fields) [217], especially in biomedicine [221]–[227]. Li *et al.* [227], for example, integrated rich features and word2vec embeddings into a joint BEE model with dual decomposition as an optimization method.

Liu *et al.* [239] combined the joint and pipeline approaches by decomposing BEE into a sequence of binary multi-class RE problems, with a `(trigger, argument)` structure for each relation. Such a pairwise model facilitates inference but requires additional fusion steps for predicting events with more than one argument (i.e., tuple extension), multiplies the number of examples to classify by a factor equal to the number of arguments, and cannot extract nested events.

One last research branch referred to MTL [219], [220].

These methods mainly concentrate on improving the network construction to acquire various effective features from the text.

c: DEEP LEARNING

More recent works use deep neural networks, where the previously manually defined features are replaced by automatically learned dense, hidden, and distributed representations.

Since 2014, word embeddings have revolutionized NLP, overcoming bag-of-words (BOW) modeling and treating relevant context terms even if not predetermined within rules or covered by dependency tree leaves [257]. How to design an efficient neural network architecture is the central question for DL-based EE; several strategies and variations have been employed.

Nguyen and Grishman [214] were the first researchers to apply *convolutional neural networks (CNNs)* to ED, with good performance in domain adaption. The authors improved their ED system by introducing non-consecutive convolution [215]. Chen *et al.* [69] adopted a CNN pipeline framework, capturing sentence-level clues hard to locate with dependency features and managing multiple-event sentences through dynamic multi-pooling. TEES CNN upgraded the original SVM version, incorporating biomedical word embeddings [212]. Wang *et al.* [213] proposed a CNN with multiple distributed features, including sentence topic information (i.e., latent dirichlet allocation (LDA)) and word-entity distance in the dependency tree.[13] Li *et al.* [60] utilized a parallel multi-pooling CNN and dependency-based word embeddings to capture relations among words that would be out of reach with small window sizes.[14]

Nguyen *et al.* [71] designed a *recurrent neural network (RNN)* based on bi-directional gated recurrent units (Bi-GRU) for joint EE. Nguyen and Nguyen [207] used Bi-GRU also to address entity detection, overcoming the unrealistic scenario of always relying on human annotated entities. Duan *et al.* [206] conceived a bi-directional long short-term memory (Bi-LSTM) receiving as input a concatenation of word and document embeddings to apprehend cross-document clues and topic information. Instead of using the original sentences, Li *et al.* [203] encoded the syntactic information of a whole dependency tree from leaves to root as a Bi-LSTM input for BEE on BioNLP'16. Sha *et al.* [201] introduced dependency bridges into a Bi-LSTM architecture, whereby each cell also receives information from the former syntactically related ones. Apart from dependency bridges, the syntactic dependency tree of a sentence can be directly exploited in a tree-structured RNN (rather than a strictly sequential one). Tree-LSTMs effectively coped with event arguments very far from their trigger, significantly shortening the distance between them [99], [202], [204]. Li *et al.* [99] incorporated entity information (type and function) from Gene Ontology into Tree-LSTM models with distributional representations. Zhang *et al.* [208] integrated both word-level and character-level bi-directional simple recurrent unit

(Bi-SRU) language models to jointly classify entities and triggers.

The *attention mechanism* gives another NLP pillar, a standard fixture of most SOTA language models, and a predominant method to treat each input item unequally according to its importance for a certain task. Specifically, attention filters out irrelevant noise and finds important units in the input sequence which are influential for the output, dynamically highlighting the parts of the text on which to focus. Many contributions stop at ED [153], [209], [211], but show higher effectiveness with long-term dependencies. Liu *et al.* [153] argued that argument words should be more influential than other ones. Zhao *et al.* [209] first learned document embeddings through a hierarchical attention network composed of two Bi-GRU encoders, paying word-level attention to event triggers and sentences containing events; then, they used the learned document-level information to facilitate ED with another RNN. Li *et al.* [205] proposed an attention-based Bi-GRU to extract biotope and bacteria events. Liu *et al.* [210] observed that trigger labels are nonessential to ED and that it is time-consuming for annotators to pick out the "most clearly" word from a given sentence, especially from a long sentence. To reduce manual effort, they explored ED without annotated triggers, mixing LSTM layers with an attention mechanism expected to automatically assign higher score to trigger words based on the target event type.

In addition to RNNs, *graph neural networks (GNNs)* — protagonists of a constantly growing trend — have recently been prompted as another way to encode syntactic and semantic relationships while preserving their structure. GNNs are neural models designed to directly operate on data described by graphs, capturing dependencies via message passing between nodes. In contrast to other encoders such as Tree-LSTMs, GNNs can cover more comprehensive contextual information for each node, thanks to their ability to capture features coming from all neighboring nodes, rather than just child ones. GNNs enable geometric deep learning in non-Euclidean spaces. Applying them to EE requires first identifying an intermediate graph representation. Liu *et al.* [70] devised an EE framework to jointly extract multiple events from sentences, leveraging on graph convolutional networks (GCNs) and self-attention. To gain maximum benefit from GCNs, they considered the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ resulting from the dependency tree — with a type label $K(v_i, v_j)$ alongside each edge $(v_i, v_j) \in \mathcal{E}$ — and further extended it with reversed edges (v_j, v_i) and self-loops (v_i, v_i). Other similar approaches have been explored for ED [200], also showing language-invariant properties [142], [147]. GNNs have also been at the core of several semi-supervised solutions, illustrated in Section III-D5.

Zhao *et al.* [258] modeled multiple BEE under a reinforcement learning (RL) framework, employing external biomedical KBs for representation learning of biomedical text, with experiments over GE13 and MLEE.

---

[13]Positional features are quite popular due to the following hypotheses: (i) the longer the distance between a word and the entities, the lower the probability it can be a trigger; (ii) the longer the distance between an entity and a trigger, the less chance it can be an argument.

[14]Syntactic dependency trees introduce shortcut arcs for draining information flow from a point to its target through fewer transitions, reducing the number of needed hops and connecting nonconsecutive yet distant words.

Sundry efforts only extract flat events with non-nested entity mentions (e.g., [70], [201], [207], [217], [220]) and ignore event modifiers, often preventing their application to BEE tasks.

d: HYBRID NEURAL NETWORK MODELS

Many researchers field hybrid solutions, combining different approaches to merge and balance their merits for EE with an ensemble or stacked strategy.

Some systems integrated pattern matching and ML [180], [181], [187]–[189], [195], or Markov logic networks with CNNs [190]. Liu *et al.* [188], [189], [195] faced ED and EE from biomedical text by searching for isomorphisms between graph-based patterns and sentence dependency graphs, implementing exact and approximate subgraph matching.

Other authors fused CNNs and RNNs [47], [192].

Still others adopted generative adversarial networks (GANs) [182], [193], [194], composed of two contesting neural networks: a generator and a discriminator, frequently implemented with RNNs and treated as teacher/student modules. For example, Zhang *et al.* [182] used inverse RL to focus on difficult-to-detect events within ambiguous sentences (i.e., different event types but similar contextual features), dynamically estimating the proper rewards on the difference between the labels committed by the experts and the extractor.

Wang *et al.* [196] developed a Bi-LSTM with conditional random fields (Bi-LSTM-CRF) for biomedical ED and multi-word triggers. Wei *et al.* [197] also integrated contextualized word representations to solve the problem of triggers ambiguity in different contexts. Chen [113] used a Bi-LSTM-CRF model for biomedical ED with an embedded transfer learning approach for joint training with multi-source datasets, potentially reducing the need for large amounts of target dataset annotations. Precisely, the author combined task-specific features and task-invariant features, permitting valuable knowledge transfer between source and target domains. As a result, a performance boost on MLEE was possible considering four different source domain datasets, including GE'09, EPI'11, and ID'11.

He *et al.* [198] divided biomedical ED into two stages; they built a Bi-LSTM model integrating attention mechanism (Att-Bi-LSTM) for trigger detection and a passive-aggressive online algorithm for multi-class event categorization.

Abdulkadhar *et al.* [191] employed graph-based kernels for ED and lexico-syntactic patterns for AD + AC, with biomolecular event sentences represented as dependency graphs. They adopted EnsembleSVM for the first event identification phase, combining a multiscale laplacian graph (MLG) kernel and a feature-based linear graph (FLG) kernel. By analyzing the structure of a biomedical sentence at multiple scales, MLG detected topological relationships between event nodes and subgraph associations. At the same time, FLG captured token-based, sentence-based, parsing-based, and domain-specific features.

Given the recognized value of the dependency syntax as an information source for the EE task, Zhang *et al.* [186] criticized the previous adoption of one-best results from a parser and the consequential impact of potentially incorrect dependencies. To alleviate this issue, they proposed using latent dependency forests drawn from a pre-trained parser where multiple dependency relations between words can be considered together with their probabilities. The authors designed an end-to-end solution for joint entity and event extraction with a Bi-LSTM encoder to transform a sentence into hidden representations and a GCN model to represent the forests. Such syntax-informed GCN model — also experimented on biomedical texts – automatically differentiated the salient syntactic information from noisy parsers, fine-tuned during event training.

5) SEMI-SUPERVISED LEARNING

Supervised EE models require intensive manual annotations to generalize well, are notoriously data-hungry, and may suffer from the data sparsity problem. In addition, BEE asks for still more domain-specific knowledge, repeatedly tacit. With the rapid development of modeling techniques, labeled data have gradually become the main bottleneck for putting EE into practice. Especially in open domain scenarios, how to improve the extraction performance from a small set of gold-labeled data is a pivotal challenge. Semi-supervised learning combines these event-level data (from annotated corpora or external resources) with a large amount of unlabeled data. Because semi-supervised learning exploits unlabeled data, the accuracy of classifiers is generally improved without additional annotation efforts.

a: PATTERN LEARNING

Some researchers proposed automatic pattern construction to increase recall rates [65], [167]–[175], [177]–[179]. In this regard, one of the first systems dates back to AutoSlog [167], developed in 1993. Starting from an annotated corpus, AutoSlog exploited a small set of domain-independent linguistic patterns (e.g., "passive-verb <dobj>") to automatically construct domain-specific dictionaries of event patterns (e.g., "took <paracetamol>") and triggers (e.g., "took"). The matching process first used the silver trigger word dictionary to spot candidate event sentences and then performed POS tagging to associate the syntax features with comparable event templates. To avoid the dependence on domain-specific tags, AutoSlog-TS [168] put a new spin on the original system by applying it to an untagged but pre-classified training corpus (i.e., texts manually labeled as either relevant or irrelevant). Yangarber *et al.* [170], [171] proposed to accomplish generation only from a small set of seed patterns, bypassing the need for a sub-corpus of relevant passages. Huang and Riloff [216] used AutoSlog to produce lexico-syntactic pattern features for generating role filler hypotheses based on the local context surrounding a noun phrase. In the biomedical sphere, Bui *et al.* [65], [177] mapped annotated events from the training data into structured representations to

decompose nested and complex events into syntactic layers, thus facilitating the automatic extraction of rules via predefined templates. Le Minh *et al.* [178] extracted heuristic rules to identify candidate triggers from training corpora, searching for tokens with appropriate POS tags near protein mentions (e.g., ''NN/NNS + of + PROTEIN'', ''VBN + PROTEIN''). After a dictionary-based ED phase, they spotted arguments and their role by creating additional patterns from textualized dependency graphs. Kilicoglu and Bergler [179] operated in a similar vein, constructing an event trigger dictionary and identifying biomedical participants through the extraction and ranking of dependency path patterns from training data. Other works [172], [175] also operated with WordNet to consider semantic affinity (e.g., the word ''sniper'' is likely to be the *agent* of a *Die* event). For example, Patwardhan and Riloff [175] ranked the patterns extracted by AutoSlog-TS based on their tendency to extract noun phrases semantically related to a specific set of event roles (e.g., ''cases of <np>'', ''spread of <np>'', ''outbreak of <np>'' for ''Disease''). By doing so, they distinguished primary patterns from secondary patterns, applying them selectively to relevant sentences (also from biomedical abstracts). As for advanced OIE systems appliable to weak ODEE, Mausam *et al.* [176] proposed OLLIE (then evolved to Open IE 5.1[15]) following the idea of bootstrap learning based on dependency parse paths. They analyzed the context to include attribution and clausal modifiers, also identifying binary relationships mediated by nouns and adjectives (instead of verbs). According to a recent survey [250], OLLIE has been the only OIE contribution evaluated on biomedical text.

b: DATA EXPANSION WITHOUT EXTERNAL RELATIONAL KNOWLEDGE

These solutions enlarge the training set by taking full advantage of the available CDEE labels for the task(s) of interest, albeit few.

Self-training is a bootstrapping method based on iterative model training, reasonably used also for EE [163], [165], [166]. At each iteration, a classifier is built on the current set of labeled data and is used to predict an additional portion of unlabeled (out-of-bag) samples. Then, the most confident instances (e.g., top-k, top-k percentage, above threshold) are added to the training set for the next round. The process continues until a certain condition is satisfied (e.g., max number of iterations, no more unlabeled samples).

Another option is discovering *latent informative instances* to be added to the training pool. Active learning strategy has been experimented for EE [161]–[163], interactively querying a user or another information source to label new data points. Wang *et al.* [164] built a huge candidate set by assuming that if a word serves as a trigger in the gold dataset, all instances mentioning that word in the unlabeled dataset might also express the same event; then, they applied an

adversarial training mechanism to identify informative instances yet filtering out noise.

Munkhdalai *et al.* [163] combined self-training and active learning approaches for BEE, developing a new method to tackle the data imbalance issue by exploiting unlabeled data for self-labeling.

Boros [160] considered the generation of *morphological derivations and inflections* from dictionaries to increase training data and discriminate more trigger forms.

Some researchers have also proposed to expand training data with *new event types* compared to the reference ones in the gold dataset, achieving broad semantic coverage. Miwa *et al.* [43] produced a single system for multiple BioNLP-ST'11 tasks, thus eliminating the need to combine predictions of different models, each trained on a single corpus. They improved EventMine by learning from multiple partially overlapping corpora, presenting a heuristic aimed to fully exploit the available annotation scopes, overcoming inconsistencies, and reducing the generation of spurious negative instances for corpora-specific types. By automatically filtering and fusing annotations from various source corpora, the authors achieved comparable or higher performance than domain adaption methods, like EventMine-EasyAdapt (based on corpus-dependent and corpus-independent features) [43] and EventMine-CR (based on stacking separately trained single models, with the integration of coreference resolution) [114]. Huang *et al.* [159] applied *zero-shot transfer learning (ZTL)* to EE, only requiring a structured definition for the never-seen event types (i.e., their ontological schema) and some annotations for one or a few seen types. Observing the similarity between the event mention structures obtained with AMR and the event type schemas, they modeled EE as a generic grounding problem by mapping each AMR-based mention to its semantically closest event type in a shared cross-concept space. New event mentions were assigned to the event type most similar to the AMR projections. A branch of research also attempted to use *few-shot transfer learning (FTL)*, requiring only a few examples to activate the system in a new domain without retraining the model [154]–[158]. Initially, Nguyen *et al.* [154] exploited a large amount of training data available for some event types to improve performance in identifying event instances of another target type only specified by a small annotated set of positive examples. Peng *et al.* [155] presented a general ED framework to map an event mention to a provided event ontology, requiring only a few unlabeled event examples in plain text for each type a user wants to extract. They viewed ED as a semantic similarity problem among event mentions or event mentions and ontologies of event types. Deng *et al.* [156] proposed an event-type learner for few-shot ED with dynamic memory networks. They first learned event-type-specific meta-knowledge from only a few instances in a support set and then applied what learned to predict new event types not seen during training. To introduce a new event-based FTL baseline, they also built the FewEvent dataset by combining ACE 2005 and TAC-KBP 2017, adding external event

---

[15]Open IE 5.1 Project. https://github.com/dair-iitd/OpenIE-standalone (accessed Nov. 23, 2021).

types in specific domains, including music, film, sports, and education. Lai *et al.* [157] formulated the ED problem in a similar way, and investigated four metric-based FTL models with different encoder types (CNN, LSTM, and GCN), also testing two novel training signals based on prototype[16] clustering. Shen *et al.* [158] introduced FrameNet knowledge and further improved FTL-based ED performance in case of limited diversity of examples in the support set.

c: DISTANT SUPERVISION

Another possibility is gathering more training data with minimal supervision, drawing on external resources. A commonly used learning paradigm is distant supervision (DS) [259], first introduced for RE[17] and then adapted to the EE setting. It aims to collect annotations for unsupervised sources by applying heuristics based on samples gleaned from existing relational databases. Strictly speaking, it attempts to generate weakly labeled data by aligning event instances from KBs to plain text and then assigning labels to matched examples. On the flip side, DS risks introducing noise and false positives (higher recall and lower precision) [260]. Besides, event triggers are not given out in existing KBs, and event arguments may be scattered [87].

Linguistic knowledge is a typical reference for DS. Aguilar *et al.* [261] shown that all events, relations, and attributes represented by ACE, ERE and TAC-KBP standards can be mapped to FrameNet[18] representations by means of some adjustments. Borrowing from this finding, Liu *et al.* [152], [153] operated with FrameNet to improve ED on the observation that many frames actually express events and share highly similar structures with ACE ones (i.e., *frame_lexical_unit↔event_trigger*, *frame_elements↔event_arguments*). Inspired by these contributions, Li *et al.* [88] leveraged FrameNet to construct fine-grained and wide-coverage hierarchical event schemas, automatically defining 655 event types with 2050 argument roles. They built a joint EE model above them and reserved frame-to-frame relations as event-to-event global features, demonstrating promising results for NLP practical applications, like news summarization. To alleviate data scarcity and low coverage, Wang *et al.* [262] presented MAVEN, a general-domain dataset for ED containing 4,480 Wikipedia documents, 118,732 event mention instances, and 168 event types. To construct it, the authors facilitated the human annotation process through a preliminary automatic type labeling phase based on FrameNet and frame semantic parsers. Interestingly, no SOTA model has yet achieved promising results on MAVEN as on small datasets, which suggests that ED (not

even full EE) remains a challenging task in the real world and requires further research efforts.

Many works bring *world and domain knowledge* into play [148]–[150]. Zeng *et al.* [150] employed DS on structured KBs or tables to automatically discover event types without trigger indications and create high-quality event annotations from unlabeled texts. For evaluation purposes, they used Freebase CVTs to tag Wikipedia articles, ending up with a training dataset 14x greater than ACE 2005, constructed with hours instead of costing years of linguists' time. Rao *et al.* [149] hypothesized that an event structure is a subgraph of an AMR graph; under this assumption, they cast the EE task as a subgraph identification problem and introduced a novel AMR path-based heuristic to selectively sample the sentences obtained from DS on Biological Pathway Exchange (BioPAX) — a KB containing protein relations — and PMC articles.

Another tendency combines *linguistic and world knowledge* [87], [151]. Chen *et al.* [87] performed DS data expansion for large-scale EE using both FrameNet and Freebase, while Araki and Mitamura [151] applied DS based on WordNet and Wikipedia for open-domain ED.

d: CROSS-LINGUAL

Most EE models are trained and applied for a particular language, ignoring the massive information that could arise from multilingual clues. Furthermore, it is inconceivable to believe that we will ever have similar datasets for every language. For these reasons, cross-lingual EE is another explored avenue (especially for limited resources), which has matured on the following two intuitions.

- *Multilingual consistency*. Sentences conveying identical events but in different languages usually have similar semantic components (i.e., consistent and coherent evidence among languages).
- *Multilingual complementation*. Ambiguous expressions in a language might have non-ambiguous counterparts in other languages, providing complementary information for disambiguation.

Procedures change depending on the required starting resources. Conventionally, they presuppose *annotation transferring* via cross-lingual bridges (e.g., word/phrase alignments) or *model transferring* (i.e., the adaption of models trained in a resource-rich language for a resource-poor one) [263]. Unfortunately, publicly available gold-standard annotations for EE exist for only a few languages and are often unbalanced across languages [89].

Multilingual Event Mentions and Labels: Chen and Ji [145] came up with a co-training framework with which the labeling results (i.e., triggers and arguments) of two monolingual EE systems were projected from one language to the other by exploiting a bilingual parallel dataset, mutually enriching both the training sets. Zhu *et al.* [146] employed machine translation (MT) services to train bilingual EE models on top of bilingual word features from unaligned datasets. Hsi *et al.* [89] used a combination of

---

[16]Prototypes are representative vectors for each class in the support set. They draw a common meta-learning concept.

[17]DS for RE assumes that if two entities have a relationship in a KB, then all sentences that mention these two entities will express that relationship.

[18]FrameNet is a lexical resource established in 1997 at the International Computer Science Institute in Berkeley, California, based on a theory of meaning called frame semantics. Each frame characterizes a primary type of semantic concept, and contains some words (lexical units) that evoke it. http://framenet.icsi.berkeley.edu (accessed Nov. 23, 2021).

language-dependent and language-independent features for multilingual training, incorporating datasets of different languages. Assuming the availability of a bilingual word dictionary, Liu *et al.* [147] devised a content-dependent lexical mapping method for cross-lingual ED, obtaining good performance also in annotation-poor scenarios. First, they learned a multilingual alignment between word dictionaries, then projected each source word into a shared embedding space and ranked different neighboring translation word candidates by context-aware selective attention to find the best-suited one.

Monolingual Event Mentions and Labels: Wei *et al.* [144] adopted translated language features to increase English ED robustness with the following idea "if an English trigger word is not in the training data, the model might still recognize the trigger if it has seen its translation before". Liu *et al.* [86] achieved data expansion using MT and unsupervised trigger projection on translated sentences, tackling monolingual ambiguity with gated cross-lingual attention. Subburathinam *et al.* [142] demonstrated that event argument structural representations obtained with GCNs might be transferred between languages without any training data for the target language. However, GCNs struggle to model long-range dependencies or disconnected words in a dependency tree. To overcome this limitation, Ahmad *et al.* [143] generated structured contextual representations through a Transformer encoder with graph attention to incorporate syntactic depth and distance information in sentence embeddings. Finally, they tested cross-lingual transferability on relation and event extraction tasks, registering significant performance improvements.

### e: LARGE-SCALE PRE-TRAINED LANGUAGE MODELS

In recent years, prominent NLP and NLU advancements have been driven by developing and deploying ever-larger self-supervised pre-trained language models (LMs). Modern LMs dynamically capture the meaning of words in accordance with their context, mainly following the Transformer architecture [264] based on attention mechanisms. Pre-training on extreme-scale unlabeled text corpora has shown to capture a surprising amount of linguistic knowledge stored implicitly in model parameters [265], [266]. Adapting this "black box knowledge" to downstream tasks has allowed significant performance gains across natural language benchmarks [267]. Owing to the unbridled release of new models (e.g., SciBERT [268], BioBERT [269], ELECTRAMed [270], PubMedBERT [271], and SciFive [272] for biomedicine), some researchers have investigated their use as backbones for EE.

Rich Contextual Representations: Zhang *et al.* [137] used BERT [273] input representations within left-to-right transition-based neural models. Liu *et al.* [141] traded on BERT architecture to enhance ED generalization and robustness; they replaced triggers with placeholders to force models making predictions uniquely using context. Trieu *et al.* [44] proposed DeepEventMine, a derivative of EventMine for joint BEE, capable of predicting nested events on nested and

overlapping entities. Built on top of SciBERT, it achieved new SOTA performance in seven datasets, reporting good results even without gold entities. Expressly, starting from enumerating all possible text spans of a sentence (up to a certain length), the model performed a flow of entity and trigger detection, role detection, event and modification detection in an end-to-end-manner through custom layers. In this regard, we observe that span-based MTL extraction methods have gained much popularity in the last years because of their ability to detect nested entity mentions. Wadden *et al.* [100] combined cross-sentence BERT embeddings and span-based graph propagation to incorporate global context in a general MTL framework, called DYGIE++. Ramponi *et al.* [140] cast BEE as a sequence labeling task (i.e., multi-task word-level tagging), introducing an efficient and accurate end-to-end system with BioBERT as encoder. Huang *et al.* [139] adopted the Unified Medical Language System (UMLS) — a KB for biomedical terminology and standards — to create hierarchical sentence-level knowledge graphs (KGs), pursuing assisting BEE under non-indicative triggers and low-resource scenarios. Intuitively, biomedical background knowledge provides clues for better identifying triggers and their corresponding arguments through factual relational reasoning, especially when the expressions used in the text are ambiguous and not sufficient to clearly recognize an event or classify a semantic role for a particular entity or sub-event. The framework supplemented SciBERT by a new GNN called GEANet to encode complex domain knowledge from sentence graphs, also integrating edge features into the attention mechanism for message propagation. Pointedly, the authors first constructed a global KG for the entire GE'11 corpus as the minimum spanning tree connecting the mentioned UMLS concepts,[19] then derived a sentence-level KG for each sentence by depth-first search. Following these footsteps, Lai *et al.* [78] proposed KECI, a span-based framework incorporating external knowledge for joint entity and relation extraction from biomedical text. Given textual input, KECI first constructs a multi-relational span graph representing its initial understanding of the text. In a span graph, each node represents a (predicted) entity mention, and each edge represents a (predicted) directed relation between two entity mentions (weighted by its probability score), where classifications are based on SciBERT encoding. KECI then uses an entity linker to form a KG containing relevant background knowledge for the detected entities. To make final predictions, it fuses the initial span graph and the KG derived from an external KB into a more refined graph, using an attention mechanism (i.e., soft alignment) between the node embeddings obtained by a GCN on each graphical source. The authors experimented KECI with UMLS, achieving leading results on two sentence-level RE datasets: BioRelEx (binding interaction detection, 33 types of entities and 3 types of relations) [275] and ADE (adverse drug event

---

[19] The extraction of UMLS biomedical entities from input documents is performed with MetaMap [274], an entity mapping tool for UMLS.
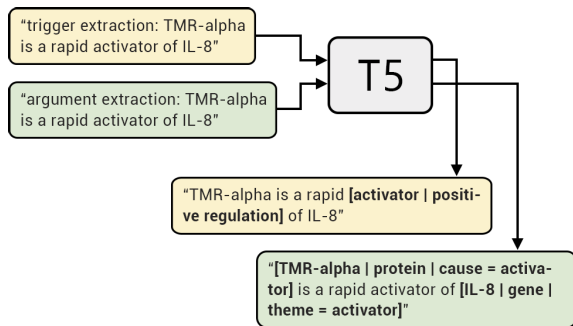
FIGURE 7. Diagram for event extraction as a text-to-text task with T5. Trigger and argument extraction subtasks share the same model, loss function, and hyperparameters.

extraction, 2 types of entities and a single relation type) [276]. Unlike [139], KECI is an MTL solution that can theoretically take an entire document as input, directly creating a KG for each input text without assuming the availability of gold-standard entity mentions at inference time. Although not yet applied to BEE, as asserted by the authors, this contribution is closely linked to the research area covered in this survey. So, we report it to inspire future event-based knowledge-enhanced models.

Event Mention Generation: Yang *et al.* [72] employed BERT for generating a controllable number of labeled event mentions by argument replacement and adjunct tokens rewriting, obviating the need for task-specific architectures or external resources. From a farther standpoint, transformers have been largely used to rewrite sentences while preserving their meaning [277], [278].

EE as Text-to-Text: Paolini *et al.* [136] presented a unified text-to-text approach for several structured prediction tasks, designing augmented natural languages to encode annotations as part of the input or output, as exemplified in Figure 7. By streamlining predictions to translations and using natural-language-like labels, they fine-tuned T5 [279] — a unified text-to-text transformer-based model — for all the tasks (prefixing the input with task ids) and offered additional benefits in low-data regimes. Although not initially intended for events, some interesting works have been recently released considering semi-structured data on one side. Bevilacqua *et al.* [280] proposed a symmetric approach for both text-to-AMR and AMR-to-text. In particular, they fine-tuned BART [281] — a transformer-based denoising autoencoder — to reconstruct English sentences from AMR *linearized graphs* and vice versa. Xu *et al.* [282] presented an end-to-end data-to-text system for splitting text into a sequence of facts using SRL (roughly capturing "who did what to whom"), learning to (i) align facts with RDF predicates, (ii) order and aggregate RDF triples (i.e., sentence planning), (iii) generate text accordingly.

EE as QA: A new trend explicitly frames EE as a machine reading comprehension (MRC) problem, capturing event information through a question-answering (QA) process [85], [133], [135] (Figure 8). In this perspective, subtasks are not defined by different loss functions but by different questions. Such learning paradigm comes with three major advantages: (i) it allows us to exploit SOTA MRC models (i.e., transfer learning); (ii) it can leverage the abundant MRC datasets (i.e., cross-domain data augmentation); (iii) it enables zero-shot EE for unseen event types and roles. McCann *et al.* [283] initially brought up the idea of casting several NLP tasks as QA over a context. Du and Cardie [133] first attempted to apply these concepts to EE; they used two BERT-based QA models, one for TD + TC and the other for AD + AC (i.e., a question for each expected role and zero or more answers per question). Despite the greater naturalness and diversity than past RE solutions [62], [284], their question generation module was based on templates obtained via context-independent rule-based strategies. Wang *et al.* [135] developed a multi-turn QA framework similar to [62] for BEE, using SciBERT; they applied even more basic rules for transforming all entity and event annotations into lexically similar questions, posed in a recursive procedure to manage nested events. Actually, the most critical point to bridge MRC and EE lies in generating relevant questions describing an event schema; parallel question-event datasets are not yet available. A forward-looking solution comes from Liu *et al.* [85], who decomposed questions into two parts: (i) a short hook still focused on template-based wh-words and (ii) primary context-related information. They approached the second with a style transfer model [285] between descriptive statements (i.e., windows of words centered at a verb) and interrogative ones (wh-words excluded), trained only from Wikipedia and SQUAD unaligned samples. However, none of these works consider the effect of argument extraction order or the relations among multiple arguments. Li *et al.* [134] developed a multi-turn dialogue system with two agents to iteratively solve the EE slot filling problem by modeling argument relations. During each turn, *agent A* selects an argument role and generates a query about an event or a chosen argument, using the same question set of [133]. Then, *agent B* answers the query by identifying the correct argument or event type from the sentence. Based on the answer, *agent A* will generate a new query in the next turn of argument extraction. This iterative process enables the system to utilize the historical knowledge obtained from previous turns when extracting an argument — whose role could be otherwise difficult to settle in isolation — and update previously identified argument roles (i.e., two-way feedback). To optimize slot filling with the right argument extraction order, the authors employed an RL-based ranking aiming to maximize the likelihood of successfully inferring the argument role. Finally, they constructed a powerful event representation by combining a lexicon-based graph attention network and a BERT model.
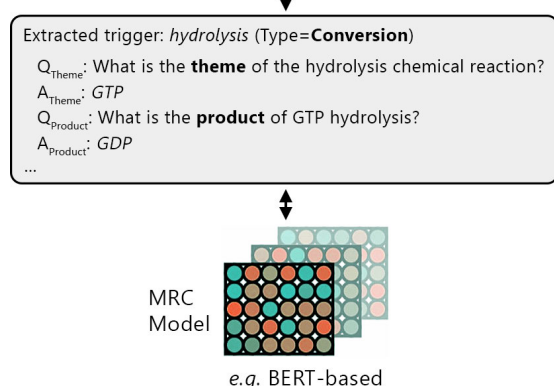
6) UNSUPERVISED LEARNING
Unsupervised learning does not rely on event labeled data and is chiefly aligned with ODEE to recognize unconstrained event types and take humans out of the loop. The lack of

TABLE 7. A biomedical document and gold event annotations from MLEE [108]. The event E4 is triggered by the word "formation" in sentence S3, while its two arguments, "NRP1" e "KDR", are distributed in sentence S2. It's effortless to note that the correct extraction of E4 requires a global context (and not a sentence-level one). Example inspired by [199].

Doc	S1	We investigated the molecular mechanisms underlying NRP1 *mediated angiogenesis* by analyzing the *association* of NRP1 and KDR.
	S2	An intracellular complex containing NRP1 and KDR was immunoprecipitated only in the presence of VEGF165.
	S3	In contrast, VEGF121 did not *support* complex *formation*.

Gold BEE labels	S1	E1<Blood_vessel_development, Trigger: angiogenesis> E2<Pos_reg, Trigger: mediated, Theme: E1, Cause: NRP1> E3<Binding, Trigger: association, Theme: NRP1, Theme: KDR>
	S2	No Event
	S3	E4<Binding, Trigger: formation, Theme: NRP1[S2], Theme: KDR[S2]> E5<Pos_reg, Trigger: support, Theme: E4, Cause: VEGF121, Polarity: Negative>

S: GAPs catalyze the hydrolysis of GTP to GDP

Extracted trigger: *hydrolysis* (Type=**Conversion**)

Q_{Theme}: What is the **theme** of the hydrolysis chemical reaction?
A_{Theme}: *GTP*
$Q_{Product}$: What is the **product** of GTP hydrolysis?
$A_{Product}$: *GDP*
...

MRC Model

e.g. BERT-based

FIGURE 8. Diagram for event extraction as a question answering task.

labels makes contributions more flexible but much more limited, tending not to pursue the ambitious goal of extracting complex, fine-grained events. Not being able to take advantage of a precise domain schema (oriented towards a well-defined set of event types) and a comprehensive dataset in this sense, unsupervised approaches are currently known for their limitations, often relaxing the definition of "event" or its level of specificity. For example, classifying event types and argument roles is not required. Moreover, unsupervised models have been applied almost entirely to news articles, social media, and heterogeneous corpora (e.g., Web, Wikipedia), overshadowing biomedical domains. While they may seem outside the scope of the survey, or a distraction from more sophisticated CDEE contributions, below, we briefly describe such research efforts that seek to operate EE without labels, hard-coded patterns, or domain restrictions (e.g., cancers, human genes, bacteria, pathways in biology).

a: META-EVENT DETECTION AND CLUSTERING
A great portion of works focuses on detecting keywords to cluster sentences/articles expressing the same event or discussing a target topic in a text stream (i.e., topic tracking) [117]–[120], [127]–[132]. Documents are typically first converted into vectorial representations (e.g., BOW [131],

TF-IDF [127], [130], lexical chains [128], multinomial distributions via Bayesian approaches [117]–[119], [129] or GANs [120]), sometimes augmented with named entities, time, and location [117], [119], [120], [129], [131], [132]. Each resulting cluster (e.g., k-means [132], agglomerative hierarchical clustering [127], sequence-based and iterative TF-IDF clustering [130], Markov clustering [131]) brings together documents describing the same high-level event (termed as meta-event), whose simplified definition depends on the previously modeled and embedded information (i.e, semantically similar words, temporal proximity, etc.).

b: EVENT SCHEMA INDUCTION
One more step concerns automatically discovering event schemas from the newly created clusters, assigning all an event type and one or more argument role names [121]–[124], [126]. Notice that the schemas thus obtained are different from the CDEE counterpart: they are unavoidably noisy and hardly named correctly. This simplified notion of schema represents a semantic synthesis that may serve as an example-based and corpus-customized annotation guideline. Huang *et al.* [126] proposed a "liberal" framework to extract events and discover their schemas from any input corpus, incorporating symbolic (e.g., AMR, FrameNet, typed dependencies) and distributional semantics. To demonstrate the approach's portability, the authors also conducted experiments on biomedical sentences, achieving 83.1% precision on trigger labeling and 78.4% on argument labeling (manual assessment). We report that some recent works have addressed the problem of automatically inducing a schema for relationships *between* events from many event instance pairs, passing the boundary of single event types. Li *et al.* [125] discovered such cross-event patterns from IE results by merging salient and semantically coherent paths into graph schemas with the adoption of path language modeling.

7) DOCUMENT-LEVEL EVENT EXTRACTION
Most research has merely focused on extracting events within the sentence scope, where a single sentence is assumed to cover complete pieces of information. Per contra,

writers usually adopt several even non-continuous sentences to articulate individual or multiple events; this also applies to many BioNLP-ST datasets [73]. In this sense, much less attention has been paid to interactions discussed across a whole document, often treating cross-sentence events as false positives. Although pre-trained bi-directional models such as BERT better capture long-distance dependencies than RNN architectures, they still have a constraint on the maximum length of the sequence, which is still below the length of many articles about events. More generally, given longer text spans, SOTA neural architectures are less effective, and task-specific self-supervision like distant supervision becomes very noisy. Existing approaches regularly decompose document-level EE into sentence-level EE, recognizing candidate event role fillers from individual sentences one by one. Overcoming this restriction is serious and stands in need of a broad text view, facing (i) argument scattering, (ii) multi-events, (iii) less explicit event relationships, (iv) long dependencies, and (v) lack of clear linguistic evidence for distant arguments linkage. Table 7 tries to summarize all these issues.

Contributions are still scarce and mainly carried out in the financial domain [115], [148], [183]. Notably, many researchers believe that target event arguments tend to be mentioned in groups of adjacent sentences, called "event regions" [148], [183], [184], [229]. Seminal works [184], [229] detected only single-sentence regions and ignored the crowding of relevant sentences. Yang *et al.* [148] assumed the closeness of the arguments and their placement around a major sentence; they considered multi-sentence event regions but only one region per document. Lately, Chen *et al.* [183] proposed to build a document graph, with nodes as candidate event arguments and edges reflecting their positional proximity or coreference relations; a graph attention network then was trained to identify multiple multi-sentence event regions and aggregate event information. Alternatively, Gupta *et al.* [286] used recurrent and recursive neural networks to classify the type of simple binary biomedical events within and across sentence boundaries, leveraging on inter-sentential dependency paths between entity pairs. Chung *et al.* [73] noted that if an entity is mentioned in a sentence, it is possible to find expressions in its adjacent sentences implying the presence of that entity. They called these expressions "context triggers" (example in Table 8). The authors presented an unsupervised method for inferring cross-sentence events by propagating intra-sentence entity information through syntactical relationships. Hence, they adopted linguistically motivated syntactic rules to collect salient context triggers from large-scale unlabeled text and extract bacteria events. Zhao *et al.* [199] then suggested an end-to-end document-level framework for BEE. The authors introduced novel Hypergraph Aggregation Neural Network (HANN) layers. First, they derived a distributed representation for each sentence through a GCN on top of a dependency parsing tree (i.e., local context). Second, they utilized a hypergraph aggregation procedure, viewing each sentence as a hyperedge in a document hypergraph

TABLE 8. Example of context trigger signaling the presence of an entity mention in an adjacent sentence, taken from [73]. Considering the meaning of the second sentence, "carriage" involves some bacteria entity mentioned in the around context, which, in this case, would be the closest one, "K. kingae". Since "carriage" and "pharyngeal" have a syntactically close relationship, it is possible to derive the cross-sentence event *Lives_In*(*K.kingae*, *pharyngeal*).

S1	None of the colonized children experienced an invasive $[K.kingae]_{BACTERIA}$ infection.
S2	The prevalence of $[pharyngeal]_{LOCATION}$ carriage among surgical patients was 8.0%.

(i.e., global context capturing attention-based relationships among sentences). By stacking multiple HANN layers, they modeled fine-grained interactions between local and global contexts; trigger and argument extraction were jointly performed on final concept representations. Instead of directly filling a document event table, Zheng *et al.* [115] generated an entity-based DAG through sequential and tractable path-expanding sub-tasks, encoding entities with document-level contexts. Du and Cardie [185] proposed an end-to-end multi-granularity reader to dynamically incorporate paragraph- and sentence-level contextualized representations. More recently, Huang and Peng [138] extended DYGIE++ — an end-to-end span-based IE framework cited in Section III-D5 — to jointly learn entity classification, entity coreference, event extraction, and event coreference. The work efficiently captured cross-event dependencies for better EE performance at the document scope.

8) PERFORMANCE COMPARISON

As reported in Section III-C, CDEE has different task definitions and evaluation criteria depending on the annotated corpus to which it refers. Far from it, ODEE does not have standardized benchmarks, and typically undergoes human evaluation, cluster metrics, contingency tables, R-precision, or normalized mutual information. Ergo, it is not easy to fairly compare all systems.

Table 9 summarizes the performance achieved by the most representative solutions addressing CDEE in the biomedical domain, taking the most popular datasets encountered in this review as references. For the reasons outlined in Section III-D6, it is not yet possible to evaluate ODEE systems on datasets designed for CDEE, with complex typed interactions and event modifiers, on which we focus.

Most of the DL models obtained superior performance compared to traditional feature-based attempts. SOTA BEE methods hover around 61.54% F-score on average over multiple shared tasks (+8% compared to the original BioNLP winners). At the time of writing, DeepEventMine [44] significantly outperforms competing systems. On the one hand, this result reflects the progress made possible by today's LMs in joint architectures; on the other, it makes evident the complexity of biomedical events compared to events more in general. For example, SOTA contributions on ACE 2005 are close to 70% F-score. Being beyond the scope of this survey,

TABLE 9. Most significant biomedical event extraction results on popular corpora, evaluated on approximate span and recursive criteria. Performance is reported in Precision, Recall, and F-score, measured on test sets. Shared task winning results are indicated with †. The highest F-score for each corpus is shown in bold.

Corpus	P (%)	R (%)	F (%)	System/Team	System Details	
					Approach	Subtask Man.
GE11	56.38	49.19	52.52	STSS [163]	Self-training + active learning	pipeline
	68.18	43.18	52.84	BioSem [65]	Pattern expansion	—
	57.65	49.56	53.30	TEES-SVM [58]	SVM	pipeline
	63.19	49.72	55.65	EventMine EasyAdapt [43]	SVM	pipeline
	64.75	49.41	56.04 †	FAUST [225]	Stacked structured predictions	joint
	64.92	51.25	57.28	EventMine Multiple [43]	SVM	pipeline
	63.48	53.35	57.98	EventMine-CR [114]	SVM	pipeline
	63.61	53.42	58.07	BioMLN [190]	MLN and SVM	joint
	69.45	49.94	58.10	TEES-CNN [212]	CNN	pipeline
	64.51	53.07	58.23	Yu et al. [204]	BiLSTM + TreeLSTM	joint
	59.33	57.37	58.33	QA with BERT [135]	EE as QA	joint
	52.14	67.01	58.65	KB-driven Tree-LSTM [99]	Tree-LSTM	joint
	64.61	56.11	60.06	GEANet-SciBERT [139]	SciBERT + GNN	joint
	69.72	53.00	60.22	BEESL [140]	EE as sequence labeling	joint
	71.73	53.21	61.10	HANN [199]	Stacked HANN layers	joint
	63.45	60.12	61.74	Abdulkadhar et al. [191]	EnsembleSVM + graph kernels	pipeline
	76.28	55.06	**63.96**	DeepEventMine [44]	SciBERT + custom layers	joint
EPI11	44.51	28.88	35.03	FAUST [225]	Stacked structured predictions	joint
	55.39	49.06	52.03	EventMine-CR [114]	SVM	pipeline
	65.70	44.70	53.21	EventMine EasyAdapt [43]	SVM	pipeline
	53.98	52.69	53.33 †	TEES-SVM [58]	SVM	pipeline
	54.42	54.28	54.35	EventMine Multiple [43]	SVM	pipeline
	64.93	50.00	56.50	TEES-CNN [212]	CNN	pipeline
	78.34	56.39	**65.57**	DeepEventMine [44]	SciBERT + custom layers	joint
ID11	48.62	37.85	42.57	TEES-SVM [58]	SVM	pipeline
	65.97	48.03	55.59 †	FAUST [225]	Stacked structured predictions	joint
	66.48	50.66	57.50	TEES-CNN [212]	CNN	pipeline
	54.97	60.55	57.63	EventMine-CR [114]	SVM	pipeline
	54.82	64.01	59.06	EventMine Multiple [43]	SVM	pipeline
	61.33	58.96	60.12	EventMine EasyAdapt [43]	SVM	pipeline
	68.51	55.99	**61.62**	DeepEventMine [44]	SciBERT + custom layers	joint
MLEE	51.48	49.36	50.40	FAUST [225]	Stacked structured predictions	joint
	58.36	46.56	51.80	Yu et al. [204]	BiLSTM + Tree-LSTM	joint
	51.23	55.23	53.15	BioMLN [190]	MLN and SVM	joint
	51.11	55.73	53.32	EventMine EasyAdapt [43]	SVM	pipeline
	60.65	45.58	53.95	TEES-CNN [212]	CNN	pipeline
	60.77	54.03	57.20	KB-driven Tree-LSTM [99]	Tree-LSTM	joint
	60.56	56.23	58.31	MultiRep-CNN [213]	CNN	pipeline
	67.23	53.61	59.65	PMCNN [60]	CNN	pipeline
	63.91	56.08	59.74	HANN [199]	Stacked HANN layers	joint
	66.08	55.02	60.05	Zhu and Zheng [47]	CNN + RNN	joint
	69.91	55.49	**61.87**	DeepEventMine [44]	SciBERT + custom layers	joint
GE13	62.83	42.47	50.68	BioSem [177]	Pattern expansion	—
	56.32	46.17	50.74 †	TEES-SVM [235]	SVM	pipeline
	58.03	45.44	50.97	EVEX [240]	TEES SVM + re-ranking	pipeline
	57.66	45.96	51.15	HDS4NLP [239]	SVM pairwise extraction	joint
	65.78	44.38	53.00	TEES-CNN [212]	CNN	pipeline
	59.71	47.96	53.19	Li et al. [227]	Dual decomposition + word2vec	joint
	59.24	48.95	53.61	BioMLN [190]	MLN and SVM	joint
	67.08	49.14	56.72	DeepEventMine [44]	SciBERT + custom layers	joint
	60.55	56.23	**58.30**	Abdulkadhar et al. [191]	EnsembleSVM + graph kernels	pipeline
CG13	64.17	48.76	55.41	TEES-SVM [235]	SVM	pipeline
	55.82	48.83	52.09 †	EventMine [46]	SVM	pipeline
	66.55	50.77	57.60	TEES-CNN [212]	CNN	pipeline
	65.81	51.91	58.04	Zhu and Zheng [47]	CNN + RNN	joint
	72.23	53.92	**61.74**	DeepEventMine [44]	SciBERT + custom layers	joint
PC13	48.74	47.85	48.29	QA with BERT [135]	EE as QA	joint
	55.78	47.15	51.10 †	TEES-SVM [235]	SVM	pipeline
	53.48	52.23	52.84	EventMine [46]	SVM	pipeline
	62.16	50.34	55.62	TEES-CNN [212]	CNN	pipeline
	61.95	50.65	55.73	Zhu and Zheng [47]	CNN + RNN	joint
	68.13	50.07	**57.72**	DeepEventMine [44]	SciBERT + custom layers	joint
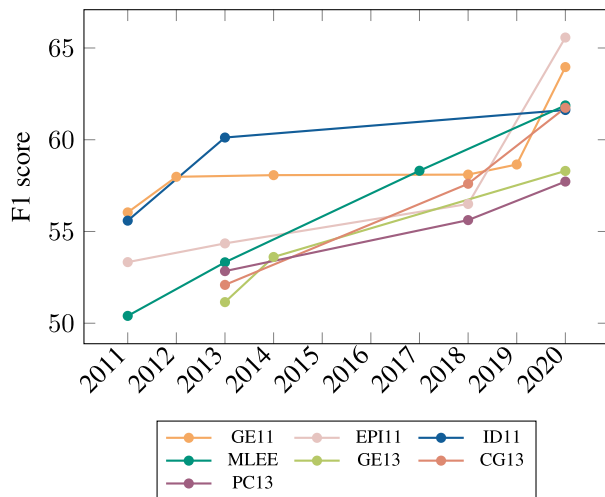
FIGURE 9. Biomedical event extraction performance on popular corpora test sets over time.

we refer the reader to [26] for further information on ACE performance.

Staying in the biomedical area, recall (i.e., sensitivity) is the measure that most limits the pursuit of high F-scores, to the detriment of a more than sufficient precision (i.e., specificity). On SOTA settings, we notice an average difference of 15.82 points between them. Highly detailed annotations (i.e., number of event and entity types, nesting percentage, etc.) adversely affect the performance. Comparably, full-texts are a more challenging EE source than abstracts due to the greater language complexity in the articles' main body [8]. Event type complexity should also be considered: while ≈80% F-scores have been achieved to extract simple events like *Activation*, accuracy can dramatically drop for more intricate events such as *Catalysis* and *Regulation* [44].

Overall, the seven analyzed datasets focus on very different subareas of the biomedical domain. Despite this, recent end-to-end DL works such as DeepEventMine [44], Zhu and Zheng [47], and HANN [199] were able to push the SOTA results of several datasets, indicating great generalization capabilities.

Although BEE has received considerable research attention in the last decade, F1 growth has been rather stagnating until the last couple of years (Figure 9). While more and more advanced models have been proposed continuously, benchmark datasets have instead upgraded slowly.

IV. CURRENT TRENDS AND BIOMEDICAL EVENT EXTRACTION RESEARCH GAPS

Most of the latest EE contributions tackle sub-tasks in an end-to-end manner without ever breaking backpropagation. MTL approaches have become popular since 2019, with a now overt tendency towards integrating at least the NER task. Such integration can further automate the discovery and curation of novel knowledge. Even so, many works still focus solely on ED, a research thread motivated by low average recall scores.

For adaptation to new domains, transfer learning carries far more expectations than ODEE. In biomedicine, the latter appears to have been tried and abandoned due to its severe limitations. The new possibilities offered by NLP break-throughs and the exploitation of the numerous existing KBs make it easier and easier to use neural models on domains not seen during training. Feasible fine-tuning operations, few-shot solutions, or even zero-shot settings achieve this goal.

There is an increasing importance of graph-based approaches for EE, prominently marked also in biomedical NLP [287]. Graph representations bridge semantics and syntax without sacrificing their structural properties, unlocking a deep understanding of texts like scientific papers and clinical narratives. Research efforts have been mainly devoted to dependency graphs from accurate parsers, moving from graphical patterns to hand-crafted features for traditional ML, from DL architectures on non-sequential input data to modern GNNs. Since 2018, we have observed a progressively wider adoption of GNNs, especially in conjunction with pre-trained LMs. In the future, we anticipate an accentuation of this trend, with a focus on external graphical sources like KGs and domain-specific ontologies.

As evidenced by our review, the reader should also be aware that only a few EE works make decisions at a document level, despite this requirement gaining more attention. The latest research is also starting to explore non-atomic EE [125], avoiding the overly simplistic assumption that events are atomic occurrences and giving more consideration to dependencies among events, fundamental for multiple extractions.

Putting aside their potential or not direct applicability to biology, it is nevertheless necessary to recognize that a much greater number of approaches have been tested in the news field (as previously depicted in Figure 5). Many of these methods have been unexplored or under-explored in the biomedical context, leaving the door open for future opportunities. Going into the details, although joint mining has become a common practice in BEE, only a couple of works [44], [186] have gone so far as to do MTL, in any case limiting themselves to the integration of entity recognition only. Instead, as we know, some span-based IE frameworks [100], [138] have been recently evaluated on ACE 2005 to integrate also entity and event coreference optimizations, achieving both high performance and computational efficiency. RL could help trigger and argument detection of biomedical events; however, we encountered only one work on this topic [258]. Document-level BEE has not yet been sufficiently covered in the literature, but HANN's results [199] look promising. Importantly, recent semi-supervised EE methods are by and large poorly applied to biomedicine, where many hints come from other domains. Some techniques not yet tested for BEE include training data expansion strategies for capturing new event types, distant supervision based on linguistic and world knowledge, cross-lingual systems, and innovative approaches designed for casting EE to other tasks over which leverage SOTA LMs. For instance, text-to-text solutions [136] could be easily adapted

to BEE. Despite the availability of various pre-trained LMs on biomedical text, these have not yet been extensively used and compared with each other as for other NLP tasks in the general domain, often relying exclusively on SciBERT or BioBERT. Transfer learning across biomedical subareas for BEE has so far been achieved through a combination of different source datasets [43], [113] or task-specific models, without investigating zero- or few-shot solutions. Even if unsupervised EE techniques are not strongly applicable to the biomedical domain, schema induction contributions appear scarce, albeit valuable for extending the semantic coverage of existing BEE systems. There are also a number of hints derivable from RE contributions not yet carried in the event world, like MTL span-based systems integrating external KBs via GCNs [78], graph-text symmetric conversions [280], and data-to-text systems with fact ordering subtasks [282].

V. APPLICATIONS AND FUTURE RESEARCH DIRECTIONS

This section tries to briefly crystallize the possible uses of events after their extraction and some of the most promising research directions for NLU purposes.

Human languages frequently involve the description of real-world events. Remarkably, EE can produce valuable structured information to facilitate a variety of downstream tasks and practical applications in diverse domains. For example, beyond biology, news, and finance, event understanding widely finds important use cases in different scientific disciplines like chemistry [288], [289].

BEE can help biomedical scientists to do research conveniently and provide inspiration and basis for the diagnosis, prevention, treatment, and new drug research. Event models have strengthened *biomedical text mining* [290], [291], *knowledge base population and enrichment* [64], [292]–[294], *biomedical literature curation* [295], *pathway curation* [296], and *biomedical literature-based knowledge discovery* [297], [298]. EE permits the construction of knowledge networks, above which execute high-level semantic queries and infer new associations. *Biological interaction networks* narrow down the search space when exploring millions of online biomedical publications [255]. Sequences of clinical event sets can improve the identification of disease phenotypes for *diagnosis prediction* [299]. In addition, events boost *reading comprehension* tasks requiring complex reasoning over paragraphs describing biological processes, mapping questions to formal queries executed against predicted rich structured representations [300]. Event-based QA contributions also come from the news domain [301].

More broadly, EE results have been applied for extractive/ abstractive single- and multi-document *summarization* (SDS and MDS) [88], [302]–[304], condensing the gist of scientific papers or news articles. Structured IE has been deemed essential for newly constructing the first biomedical MDS dataset [305], evaluating information content preservation against weaker metrics such as ROUGE. Events have been used to support *semantic search* [306] and *information retrieval* [256], seeing them as digital document metadata.

A user can conduct queries by interaction examples, wholly or partially specifying the elements of an event type. Furthermore, textual event understanding has been applied for *predicting people's intents and reactions* [307]. Learning causal relations of events is beneficial for *narrative comprehension* to foretell what happens next in a story [308] or for *stock market predictions* [309].

Nonetheless, the fast development of research areas like DL and KGs have rekindled enthusiasm towards EE, previously tempered by the intrinsic complexity of the task. Today, many NLP experts agree that event-centric NLU impacts and will impact a broad spectrum of knowledge-driven AI challenges [310]. In this scenario, a big step on the event ground may come from the following open problems.

1) *More extensive LM usage*. The adoption of large-scale pre-trained LMs for EE is still at the beginning, with few contributions showing their great potential, especially in terms of representation learning. Many models and architectures have not yet been tested in the biomedical domain. The newborn transformer efficiency trend [311] would allow handling longer inputs, useful for document-level extraction.

2) *Integration of external knowledge for EE*. Many recent models rely on distributional representations and neglect to incorporate external structured knowledge for providing rich factual information to support IE. Given the continual expansion of biomedical literature, the high number of mentionable entities, the complexity of the facts expressed, and the need for reliable systems, this is a vital need in biology. In this sense, Lai *et al.* [78] showed that a SciBERT model — even if pre-trained on 1.14M scientific papers — still incorrectly predicts the type of several entities, like "DNA" for the term "UIM" instead of "Protein Motif". Surprisingly, the high amount of world knowledge coming from biological databases has not yet been largely used to improve the performance of BEE tasks. Current contributions focus on improving entity embedding with additional semantic information, labeling new biomedical text with distant supervision, and building sentence-level graphs augmented with background knowledge as bases for applying GNNs. Several modern techniques remain uncovered, such as graph verbalization to generate new artificial training documents [312] and retrieval-based LMs to attend external documents helpful for predictions during pre-training, fine-tuning, and inference [313]. It is reasonable to expect that EE models that will not go beyond their fixed set of parameters will likely fall behind in the upcoming years, mainly if applied to biology.

3) *Multi-domain EE*. Most EE systems are trained on a single annotated corpus with restricted coverage. A dimension that should deserve further study concerns CDEE training on multiple annotated corpora, adding a subtask in the NLU pipeline to firstly select the right event schema according to the domain. A similar idea

is already applied in dialogue systems [15], [134], where a first phase aims to recognize the user's intent (e.g., flight departures query in the travel domain) and a second to search in the user's text for all the necessary information (e.g., filling predefined slots like original city, destination city, departure time, and arrival time). Combined with the cross-domain skills of NLP models [314], [315], this can allow, for example, to build a single BEE system for several biomedical sub-areas.

4) *Event aggregation*. Though texts with the same basic meaning tend to generate the same event graph, there can be numerous nuances. For example, two sentences or documents can express the same interaction in more or less detail (i.e., at different nesting levels k). By contrast, two textual spans marked with equivalent annotations may still differ, albeit referring to the same argument type. Understanding when two events express the same interaction and how to possibly aggregate them is substantially a road not yet traveled, where event embedding can be crucial. Aggregation can lead to automatically generate a formal, quantified, and semantically rich graph representation of the *concept units* mentioned in an entire corpus (rather than a naive collection of event instances, which is quite far from a true form of "knowledge"). By doing so, for example, a doctor could instantly have access to the number of times patients have complained of a particular symptom or the number of times that a drug has caused a specific side effect, normalizing many lexical and syntactic variations in the text. Alternatively to the analysis of biomedical literature and medical reports, a burgeoning research thread could come from EE on messages shared by patients and caregivers within social communities [316]–[319]. Events may be a driver of innovation in automatically capturing and measuring "the voice of patients", as a set of real-world problems, experiences, and feelings often not reported to physicians and systematically lost.

5) *Other event-based tasks*. Much attention has been paid to EE, but other tasks need to be deepened, such as inter-event relation extraction (e.g., coreferential, causal, and temporal relations) and event embedding (to find semantic and structural similarities between different interactions mapped in low dimensional vector spaces). Deep graph kernels have achieved impressive results in quantifying event similarity within unsupervised settings [320]; exciting advancements could also come from deep metric learning [321].

6) *Hypergraphs*. Some EE models can be generalized and applied to other NLP tasks, aiming to represent and extract any hypergraphs from the text. A thread also envisaged by [44].

7) *Integrating event knowledge into LMs*. Currently, it is difficult to determine what knowledge is stored in a LM and where. Interestingly, as these models have become larger (and trained on more data), their benchmark performance has continued to grow despite limited conceptual improvements, leaving open questions about the origin of these remarkable generalization properties. Recent research activity [322] has hypothesized that many of these performance gains could derive from *relational facts* (in addition to linguistic knowledge), learned during training and then exploited at evaluation time. Petroni *et al.* [323] have discovered that accurate commonsense factual knowledge (i.e., triples in subject-relation-object $\{s, r, o\}$ format) can be recovered surprisingly well from pre-trained LMs, without fine-tuning. Thereupon, the paradigm of LMs as KBs has emerged, leading to the optimistic claim that LMs already fully encode commonsense and remove the need for structured knowledge resources [324]. On the contrary, Hwang *et al.* [249] have proved the utility and importance of KGs as *transfer resources*, finding out that KG-adapted LMs learn to express relational knowledge much more precisely than zero-shot LMs. Particularly, they fine-tuned LMs on the task of predicting o given s and r, referring to $ATOMIC^{20}_{20}$ with the idea of guessing learned knowledge through tail generation. By training on KGs, generation accuracy skyrocketed: for instance, a BART-based *knowledge model* beat GPT-3 [325] by $\approx$12 points (human evaluation) and 430x fewer parameters. He *et al.* [326] injected medical knowledge into pre-trained LMs, outperforming BERT-Base on several medical NLP tasks. They transposed the famous linguistic distributional hypothesis "you shall know a word by the company it keeps" to the KG world: "you shall know an entity by the relationships it involves". To capture knowledge in a more modular and interpretable way, Guu *et al.* [313] have proposed REALM, a retrieval-based LM. Instead of storing all the knowledge in network parameters, REALM augments LM pre-training with a textual knowledge retriever learned above a large corpus (like Wikipedia), asking the model to decide what knowledge to retrieve and use during inference. Agarwal *et al.* [312] have explored converting KGs to synthetic natural language sentences for corpora augmentation, integrating them into LM pre-training without architectural changes and overcoming structural differences. To that end, they combined REALM with a synthetic corpus obtained by *verbalizing* the entire Wikidata KG, showing substantial benefits over QA datasets. So, while one research thread wants to extract structured knowledge from the text automatically, the other looks to inject structured knowledge sources into LMs. In this context, events offer a great opportunity to combine symbolic and subsymbolic approaches. Large event representations can be used as input sources for DL algorithms, bridging two less and less distant worlds. Event-based KGs could provide a semantic and symbolic layer to support NLP and NLU tasks generally tackled with brute force approaches on text. Event knowledge injection would thus constitute a foundational contribution. In biology, it would give the

possibility to a neural network to utilize all the biomedical interactions extracted from the literature (annotated or predicted, i.e., gold or silver) as background knowledge for improving performance on any biomedical NLP task. This integration is also expected to bolster factuality, reduce toxicity and hallucinations [327]. The quest remains *to what extent and in what way can event-driven modeling be superior to traditional language modeling?*

8) *Explainability*. Events can be helpful not only to facilitate task resolution but also to provide an effective way to interpret the black-box answers given by neural models. In fact, event knowledge is well interpretable, unlike matrices of real numbers learned within neural layers. Conceptual graph structures are known to be good candidates for a general explanation representation in XAI [328], [329]. Therefore, using events within NLP systems is a promising direction to foster their interpretability or explainability, also working on metrics to gauge these properties objectively. In the biomedical domain, more than others, the reliability of a system is a key requirement that cannot be ignored.

9) *Neural networks with event-based memories*. A big step forward on the road to high-performance, explainability, and knowledge injection could originate from incorporating human-like memories into neural networks [330]–[333], providing the latter with the ability to perform *Read* and *Write* operations at their or our convenience. The storage and node/edge editing of concept unit graphs like events could expose the knowledge learned by a model in a human-comprehensible way. In this context, we observe a partial transfer of concepts from the past, appropriately revised, where models based on algebraic or probabilistic methods with latent spaces applied to the text [334], [335] continue to have a value.

10) *Event algebra*. As relational algebra exists, it would be interesting to define an event algebra, creating a closed mathematical structure with operations between events (e.g., selection, join, aggregation) that enjoy usual algebraic properties. Thanks to its greater expressiveness, a theoretical study of the event paradigm could allow us to reach a parallel between a new algebra and a language of events, similar to that already occurred between relational algebra and the SQL language, with an epochal impact in data representation and manipulation.

VI. CONCLUSION

This paper has conducted a comprehensive survey on event extraction (EE) from text, which has been identified as one of the most important research themes for natural language understanding and biomedical literature analysis. Our study has first clarified different concepts underlying events and showed a detailed summary of the available annotated corpora for closed-domain biomedical EE (BEE). These conceptual remarks have served as a solid baseline for a systematic review of the core contributions in the field. To this end,

we have provided a global taxonomy of the works reported by the community, classifying them under multiple criteria; namely, schema dependency, subtask management, analysis granularity, exploited features, and approach. Next, we have compared the performance achieved by the most significant BEE models, aiding researchers and practitioners in making well-informed decisions. Subsequently, we have analyzed current literature trends and identified the gaps of EE research in the biomedical field compared to that conducted in other domains, giving insights for applying strategies not yet explored. Finally, we have progressed the discussion on applications and promising research directions. Our reflections about the future of EE corroborate its potential to automatically extract and represent the valuable knowledge encoded within text.

AUTHORS' CONTRIBUTIONS

All authors contributed to the design and production of the final manuscript, providing critical feedback and helping to shape the research. Giacomo Frisoni conducted the review and wrote the manuscript. Gianluca Moro and Antonella Carbonaro supervised all steps of the work and revised each section.

REFERENCES

[1] E. Landhuis, "Scientific literature: Information overload," *Nature*, vol. 535, no. 7612, pp. 457–458, Jul. 2016.

[2] M. A. Valenzuela-Escárcega, Ö. Babur, G. Hahn-Powell, D. Bell, T. Hicks, E. Noriega-Atala, X. Wang, M. Surdeanu, E. Demir, and C. T. Morrison, "Large-scale automated machine reading discovers new cancer-driving mechanisms," *Database J. Biol. Databases Curation*, vol. 2018, p. bay098, Jan. 2018.

[3] M. S. Simpson and D. Demner-Fushman, "Biomedical text mining: A survey of recent progress," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds. Boston, MA, USA: Springer, 2012, pp. 465–517.

[4] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction," *Frontiers Cell Develop. Biol.*, vol. 8, p. 673, Aug. 2020.

[5] L. Kong, C. Winestock, and P. Bhatia, "Zero-shot medical entity retrieval without annotation: Learning from rich knowledge graph semantics," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Assoc. Comput. Linguistics, Aug. 2021, pp. 2401–2405. [Online]. Available: https://aclanthology.org/2021.findings-acl.212, doi: 10.18653/v1/2021.findings-acl.212.

[6] J. G. Zheng, D. Howsmon, B. Zhang, J. Hahn, D. McGuinness, J. Hendler, and H. Ji, "Entity linking for biomedical literature," *BMC Med. Informat. Decis. Making*, vol. 15, no. S-1, p. S4, 2015.

[7] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "GENETAG: A tagged corpus for gene/protein named entity recognition," *BMC Bioinf.*, vol. 6, no. S-1, p. S3 2005.

[8] J. A. Vanegas, S. Matos, F. A. González, and J. L. Oliveira, "An overview of biomolecular event extraction from scientific documents," *Comput. Math. Methods Med.*, vol. 2015, pp. 571381:1–571381:19, Aug. 2015.

[9] C. Wei, K. Lee, R. Leaman, and Z. Lu, "Biomedical mention disambiguation using a deep learning approach," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform. (BCB)*, Niagara Falls, NY, USA, X. M. Shi, M. Buck, J. Ma, and P. Veltri, Eds., Sep. 2019, pp. 307–313.

[10] Q. Chen, K. Lee, S. Yan, S. Kim, C.-H. Wei, and Z. Lu, "BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale," *PLOS Comput. Biol.*, vol. 16, no. 4, Apr. 2020, Art. no. e1007617.

[11] H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of gene-disease relations from medline using domain dictionaries and machine learning," in *Proc. Pacific Symp. Biocomputing*, Maui, HI, USA, R. B. Altman, T. Murray, T. E. Klein, A. K. Dunker, and L. Hunter, Eds., Jan. 2006, pp. 4–15.

[12] S. K. Sahu and A. Anand, "Drug-drug interaction extraction from biomedical texts using long short-term memory network," *J. Biomed. Inform.*, vol. 86, pp. 15–24, Oct. 2018.

[13] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, and L. Yang, "A hybrid model based on neural networks for biomedical relation extraction," *J. Biomed. Informat.*, vol. 81, pp. 83–92, May 2018.

[14] C. Quan, Z. Luo, and S. Wang, "A hybrid deep learning model for protein–protein interactions extraction from biomedical literature," *Appl. Sci.*, vol. 10, no. 8, p. 2690, Apr. 2020.

[15] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 2, pp. 25–35, Dec. 2017.

[16] B. Y. Lin, X. Chen, J. Chen, and X. Ren, "KagNet: Knowledge-aware graph networks for commonsense reasoning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, Nov. 2019, pp. 2829–2839.

[17] L. Huang, L. Wu, and L. Wang, "Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 5094–5107.

[18] F. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong, "An overview of event extraction from text," in *Proc. Workshop Detection, Represent., Exploitation Events Semantic Web (DeRiVE)*, vol. 779, Bonn, Germany, M. van Erp, W. R. van Hage, L. Hollink, A. Jameson, and R. Troncy, Eds., Oct. 2011, pp. 48–57.

[19] F. Hogenboom, F. Frasincar, U. Kaymak, F. D. Jong, and E. Caron, "A survey of event extraction methods from text for decision support systems," *Decis. Support Syst.*, vol. 85, pp. 12–22, May 2016.

[20] V. Danilova, M. Alexandrov, and X. Blanco, "A survey of multilingual event extraction from text," in *Proc. 19th Int. Conf. Appl. Natural Lang. Inf. Syst. (NLDB)*, in Lecture Notes in Computer Science, vol. 8455, Montpellier, France, E. Métais, M. Roche, and M. Teisseire, Eds. Cham, Switzerland: Springer, Jun. 2014, pp. 85–88.

[21] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 132–164, 2015.

[22] M. Mejri and J. Akaichi, "A survey of textual event extraction from social networks," in *Proc. 1st Conf. Lang. Process. Know. Manage. (LPKM)*, vol. 1988, Kerkennah, Tunisia, L. H. Belguith and S. Mechti, Eds., Sep. 2017, pp. 1–13.

[23] S. Ananiadou, S. Pyysalo, J. Tsujii, and D. B. Kell, "Event extraction for systems biology by text mining the literature," *Trends Biotechnol.*, vol. 28, no. 7, pp. 381–390, Jul. 2010.

[24] A. R. M. A. Basher, A. S. Purdy, and I. Birol, "Event extraction from biomedical literature," *J. Bioinf. Comput. Biol.*, vol. 8, no. 1, pp. 131–146, 2015. [Online]. Available: https://www.ijcseonline.org/pub_paper/13-IJCSE-02912.pdf

[25] Z. Zhang, Y. Wu, and Z. Wang, "A survey of open domain event extraction," Dept. Comput. Sci., Univ. Illinois Urbana-Champaign, Champaign, IL, USA, Tech. Rep. 1, 2018.

[26] W. Xiang and B. Wang, "A survey of event extraction from text," *IEEE Access*, vol. 7, pp. 173111–173137, 2019.

[27] J. L. Elman and K. McRae, "A model of event knowledge," in *Proc. 39th Annu. Meeting Cogn. Sci. Soc. (CogSci)*, G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar, Eds., London, U.K., Jul. 2017, pp. 16–29.

[28] K. Liu, Y. Chen, J. Liu, X. Zuo, and J. Zhao, "Extracting event and their relations from texts: A survey on recent research progress and challenges," *AI Open*, vol. 1, pp. 22–39, Jan. 2020.

[29] D. Davidson, "The logical form of action sentences," in *The Logic of Decision and Action*, N. Rescher, Ed. Univ. Pittsburgh Press, 1967, pp. 81–95.

[30] E. Pederson and J. Bohnemeyer, "On representing events—An introduction," *Event Represent. Lang. Cognition*, pp. 1–12, 2010. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84926089842&doi=10.1017%2fCBO9780511782039.001&partnerID=40&md5=6f5e445bb60a1cb819a62a9375e46951, doi: 10.1017/CBO9780511782039.001.

[31] J. Pustejovsky, "The syntax of event structure," in *The Language of Time—A Reader*, I. Mani, J. Pustejovsky, and R. J. Gaizauskas, Eds. Oxford, U.K.: Oxford Univ. Press, 2005, pp. 33–60.

[32] C. Maienborn, *Event Semantics*, vol. 1, C. Maienborn, K. von Heusinger, and P. Portner, Eds. De Gruyter Mouton, 2011, pp. 802–829. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/9783110226614.802/html, doi: 10.1515/9783110226614.802.

[33] E. Ünal, Y. Ji, and A. Papafragou, "From event representation to linguistic meaning," *Topics Cognit. Sci.*, vol. 13, no. 1, pp. 224–242, Jan. 2021.

[34] Linguistic Data Consortium. (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*. Version 5.4.3 2005.07.01. [Online]. Available: https://www.ldc.upenn.edu/collaborations/past-projects/ace

[35] S. T. Rosen, "The syntactic representation of linguistic events," *Glot Int.*, vol. 4, no. 2, pp. 3–11, 1999.

[36] J.-D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC Bioinf.*, vol. 9, no. 1, pp. 1–25, Dec. 2008.

[37] P. F. Griffin, "The correlation of English and journalism," *English J.*, vol. 38, no. 4, pp. 189–194, 1949.

[38] J.-F. Shan, Z.-T. Liu, and W. Zhou, "Sentence similarity measure based on events and content words," in *Proc. 6th Int. Conf. Fuzzy Syst. Knowl. Discovery*, vol. 7, 2009, pp. 623–627.

[39] K. Chakma and A. Das, "A 5W1H based annotation scheme for semantic role labeling of English tweets," *Computacion y Sistemas*, vol. 22, no. 3, pp. 747–755, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055564668&doi=10.13053%2fCyS-22-3-3016&partnerID=40&md5=7f9274da310edebe42bef009412c83f5, doi: 10.13053/CyS-22-3-3016.

[40] D. Zhou, D. Zhong, and Y. He, "Biomedical relation extraction: From binary to complex," *Comput. Math. Methods Med.*, vol. 2014, pp. 298473:1–298473:18, Aug. 2014.

[41] M. Miwa, P. Thompson, I. Korkontzelos, and S. Ananiadou, "Comparable study of event extraction in newswire and biomedical domains," in *Proc. 25th Int. Conf. Comput. Linguistics*, Dublin, Ireland, J. Hajic and J. Tsujii, Eds., Aug. 2014, pp. 2270–2279.

[42] M. Miwa, P. Thompson, J. McNaught, D. B. Kell, and S. Ananiadou, "Extracting semantically enriched events from biomedical literature," *BMC Bioinf.*, vol. 13, no. 1, p. 108, Dec. 2012.

[43] M. Miwa, S. Pyysalo, T. Ohta, and S. Ananiadou, "Wide coverage biomedical event extraction using multiple partially overlapping corpora," *BMC Bioinf.*, vol. 14, no. 1, p. 175, 2013.

[44] H.-L. Trieu, T. T. Tran, K. N. A. Duong, A. Nguyen, M. Miwa, and S. Ananiadou, "DeepEventMine: End-to-end neural nested event extraction from biomedical texts," *Bioinformatics*, vol. 36, no. 19, pp. 4910–4917, Dec. 2020.

[45] S. Pyysalo, T. Ohta, R. Rak, D. E. Sullivan, C. Mao, C. Wang, B. W. S. Sobral, J. Tsujii, and S. Ananiadou, "Overview of the infectious diseases (ID) task of BioNLP shared task 2011," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Jun. 2011, pp. 26–35.

[46] M. Miwa and S. Ananiadou, "NaCTeM EventMine for BioNLP 2013 CG and PC tasks," in *Proc. BioNLP Shared Task Workshop*, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Sofia, Bulgaria, Aug. 2013, pp. 94–98.

[47] L. Zhu and H. Zheng, "Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks," *BMC Bioinf.*, vol. 21, no. 1, p. 47, Dec. 2020.

[48] H. Gui, J. Liu, F. Tao, M. Jiang, B. Norick, L. M. Kaplan, and J. Han, "Embedding learning with events in heterogeneous information networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2428–2441, Nov. 2017.

[49] C. Berge, *Hypergraphs—Combinatorics of Finite Sets* (North-Holland Mathematical Library), no. 45. Amsterdam, The Netherlands: North-Holland, 1989.

[50] K. Bollacker, C. Evans, and P. Paritosh, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 1247–1250.

[51] S. Qian and M. Amblard, "Event in compositional dynamic semantics," in *Proc. 6th Int. Conf. Log. Aspect Comput. Linguistic (LACL)*, in Lecture Notes in Artificial Intelligence, vol. 6736, S. Pogodalla and J.-P. Prost, Eds. Montpellier, France: Springer, Jun. 2011, pp. 219–234. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00601620/file/Event_in_Compositional_Dynamic_Semantics1.pdf, doi: 10.1007/978-3-642-22221-4_15.

[52] S. Pyysalo, T. Ohta, and S. Ananiadou, "Overview of the cancer genetics (CG) task of BioNLP shared task 2013," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 58–66.

[53] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, pp. 25–29, May 2000.

[54] B. Sundheim, "Overview of the fourth message understanding evaluation and conference," in *Proc. 4th Conf. Message Understand. (MUC)*, McLean, VA, USA, Jun. 1992, pp. 3–21.

[55] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proc. 16th Int. Conf. Comput. Linguistics (COLING)*. Copenhagen, Denmark: Center for Sprogteknologi, Aug. 1996, pp. 466–471.

[56] E. Riloff, "Automatically generating extraction patterns from untagged text," in *Proc. 13th Nat. Conf. Artif. Intell., 8th Innov. Appl. Artif. Intell. Conf.*, vol. 2, Portland, OR, USA, W. J. Clancey and D. S. Weld, Eds., Aug. 1996, pp. 1044–1049.

[57] S. Yan and K.-C. Wong, "Context awareness and embedding for biomedical event extraction," *Bioinformatics*, vol. 36, no. 2, pp. 637–643, Aug. 2019.

[58] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proc. BioNLP Shared Task Workshop*, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Portland, OR, USA, Jun. 2011, pp. 183–191.

[59] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: Success, failure and the future," *Briefings Bioinf.*, vol. 17, no. 1, pp. 132–144, Jan. 2016.

[60] L. Li, Y. Liu, and M. Qin, "Extracting biomedical events with parallel multi-pooling convolutional neural networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 2, pp. 599–607, Aug. 2020.

[61] A. Akbik and A. Löser, "KrakeN: N-ary facts in open information extraction," in *Proc. AKBC-WEKEX@NAACL-HLT*, 2012, pp. 52–56.

[62] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, and J. Li, "Entity-relation extraction as multi-turn question answering," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul./Aug. 2019, pp. 1340–1350.

[63] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105114155&doi=10.1109%2fTNNLS.2021.3070843&partnerID=40&md5=f1fe35912c009aeea66424df37f54e33, doi: 10.1109/TNNLS.2021.3070843.

[64] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, and T. Bogaard, "Building event-centric knowledge graphs from news," *J. Web Semantics*, vols. 37–38, pp. 132–151, Mar. 2016.

[65] Q.-C. Bui and P. M. A. Sloot, "A robust approach to extract biomedical events from literature," *Bioinformatics*, vol. 28, no. 20, pp. 2654–2661, 2012.

[66] S. Gottschalk and E. Demidova, "EventKG: A multilingual event-centric temporal knowledge graph," in *Proc. Eur. Semantic Web Conf. (ESWC)*, in Lecture Notes in Computer Science, vol. 10843, Heraklion, Greece, A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, Jun. 2018, pp. 272–287.

[67] N. Bach and S. Badaskar, "A review of relation extraction," *Literature Rev. Lang. Statist. II*, vol. 2, pp. 1–15, Nov. 2007.

[68] J. Kim, T. Ohta, K. Oda, and J. Tsujii, "From text to pathway: Corpus annotation for knowledge acquisition from biomedical literature," in *Proc. 6th Asia–Pacific Bioinf. Conf. (APBC)*, vol. 6, Kyoto, Japan, A. Brazma, S. Miyano, and T. Akutsu, Eds. London, U.K.: Imperial College Press, Jan. 2008, pp. 165–176.

[69] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process. (ACL)*, Beijing, China, Jul. 2015, pp. 167–176.

[70] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Oct./Nov. 2018, pp. 1247–1256.

[71] T. H. Nguyen, K. Cho, and R. Grishman, "Joint event extraction via recurrent neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL HLT)*, San Diego, CA, USA, K. Knight, A. Nenkova, and O. Rambow, Eds., Jun. 2016, pp. 300–309.

[72] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul./Aug. 2019, pp. 5284–5294.

[73] J.-W. Chung, W. Yang, and J. C. Park, "Unsupervised inference of implicit biomedical events using context triggers," *BMC Bioinf.*, vol. 21, no. 1, p. 29, Dec. 2020.

[74] B. Zou, G. Zhou, and Q. Zhu, "Negation and speculation target identification," in *Proc. 3rd CCF Int. Conf. Natural Lang. Process. Chin. Comput. (NLPCC)*, in Communications in Computer and Information Science, vol. 496, Shenzhen, China, C. Zong, J. Nie, D. Zhao, and Y. Feng, Eds. Berlin, Germany: Springer, Dec. 2014, pp. 34–45.

[75] R. Girju, "Negation and speculation detection," *Comput. Linguistics*, vol. 45, no. 4, pp. 819–821, 2019.

[76] A. S. Wu, B. H. Do, J. Kim, and D. L. Rubin, "Evaluation of negation and uncertainty detection and its impact on precision and recall in search," *J. Digit. Imag.*, vol. 24, no. 2, pp. 234–242, Apr. 2011.

[77] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. M. Summers, and Z. Lu, "NegBio: A high-performance tool for negation and uncertainty detection in radiology reports," *AMIA Summits Transl. Sci. Proc.*, vol. 2018, pp. 188–196, 2018.

[78] T. Lai, H. Ji, C. Zhai, and Q. H. Tran, "Joint biomedical entity and relation extraction with knowledge-enhanced collective inference," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process. (ACL/IJCNLP)*, Aug. 2021, pp. 6248–6260.

[79] M. Choi, H. Liu, W. Baumgartner, J. Zobel, and K. Verspoor, "Coreference resolution improves extraction of biological expression language statements from texts," *Database*, vol. 2016, Jan. 2016, Art. no. baw076.

[80] K. B. Cohen, A. Lanfranchi, M. J.-Y. Choi, M. Bada, W. A. Baumgartner, N. Panteleyeva, K. Verspoor, M. Palmer, and L. E. Hunter, "Coreference annotation and resolution in the Colorado richly annotated full text (CRAFT) corpus of biomedical journal articles," *BMC Bioinf.*, vol. 18, no. 1, pp. 372:1–372:14, Dec. 2017.

[81] P. Stenetorp, G. Topić, S. Pyysalo, T. Ohta, J.-D. Kim, and J. Tsujii, "BioNLP shared task 2011: Supporting resources," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, Jun. 2011, pp. 112–120.

[82] L. Hirschman, "The evolution of evaluation: Lessons from the message understanding conferences," *Comput. Speech Lang.*, vol. 12, no. 4, pp. 281–305, Oct. 1998.

[83] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "ACE program–task definitions and performance measures," in *Proc. LREC*, 2004, pp. 837–840.

[84] C. Walker, S. Strassel, J. Medero, and K. Maeda, "ACE 2005 multilingual training corpus," Linguistic Data Consortium, Philadelphia, PA, USA, 2006, p. 45, vol. 57.

[85] J. Liu, Y. Chen, K. Liu, W. Bi, and X. Liu, "Event extraction as machine reading comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 1641–1651.

[86] J. Liu, Y. Chen, K. Liu, and J. Zhao, "Event detection via gated multilingual attention mechanis," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, S. A. McIlraith and K. Q. Weinberger, Eds., Feb. 2018, pp. 4865–4872.

[87] Y. Chen, S. Liu, X. Zhang, K. Liu, and J. Zhao, "Automatically labeled data generation for large scale event extraction," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, BC, Canada, R. Barzilay and M. Kan, Eds., Jul./Aug. 2017, pp. 409–419.

[88] W. Li, D. Cheng, L. He, Y. Wang, and X. Jin, "Joint event extraction based on hierarchical event schemas from framenet," *IEEE Access*, vol. 7, pp. 25001–25015, 2019.

[89] A. Hsi, Y. Yang, J. G. Carbonell, and R. Xu, "Leveraging multilingual training for limited resource event extraction," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, Osaka, Japan, N. Calzolari, Y. Matsumoto, and R. Prasad, Eds., Dec. 2016, pp. 1201–1210.

[90] Z. Song, A. Bies, S. M. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma, "From light to rich ERE: Annotation of entities, relations, and events," in *Proc. 3rd Workshop EVENTS, Definition, Detection, Conf., Represent. (EVENTS@HLP-NAACL)*, Denver, CO, USA, E. H. Hovy, T. Mitamura, and M. Palmer, Eds., Jun. 2015, pp. 89–98.

[91] J. Ellis, J. Getman, N. Kuster, Z. Song, A. Bies, and S. M. Strassel, "Overview of linguistic resources for the TAC KBP 2016 evaluations: Methodologies and results," in *Proc. Text Anal. Conf. TAC*, Gaithersburg, MD, USA, Nov. 2016.

[92] J. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of BioNLP'09 shared task on event extraction," in *Proc. BioNLP Workshop Companion Volume Shared Task (BioNLP@HLT-NAACL)*, Boulder, CO, USA, Jun. 2009, pp. 1–9.

[93] J. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. L. T. Nguyen, and J. Tsujii, "Overview of BioNLP shared task 2011," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Jun. 2011, pp. 1–6.

[94] C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, "Overview of BioNLP shared task 2013," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 1–7.

[95] C. Nedellec, R. Bossy, and J. Kim, Eds., *Proc. 4th BioNLP Shared Task Workshop (BioNLP)*, Berlin, Germany, Aug. 2016. [Online]. Available: https://aclanthology.org/volumes/W16-29/

[96] J. G. Caporaso, N. Deshpande, J. L. Fink, P. E. Bourne, K. B. Cohen, and L. Hunter, "Intrinsic evaluation of text mining tools may not predict performance on realistic tasks," in *Proc. Pacific Symp. Biocomputing*, Kohala Coast, HI, USA, R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein, Eds., Jan. 2008, pp. 640–651.

[97] J. Kim, C. Nédellec, R. Bossy, and L. Deléger, Eds., in *Proc. 5th Workshop BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP*, Hong Kong, Nov. 2019. [Online]. Available: https://aclanthology.org/volumes/D19-57/

[98] H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor, "BioLemmatizer: A lemmatization tool for morphological processing of biomedical text," *J. Biomed. Semantics*, vol. 3, no. 1, p. 3, Dec. 2012.

[99] D. Li, L. Huang, H. Ji, and J. Han, "Biomedical event extraction based on knowledge-driven tree-LSTM," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 1421–1430.

[100] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 5783–5788.

[101] J. Fluck, S. Madan, S. Ansari, A. T. Kodamullil, R. Karki, M. Rastegar-Mojarad, N. L. Catlett, W. Hayes, J. Szostak, J. Hoeng, and M. Peitsch, "Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL)," *Database*, vol. 2016, Jan. 2016, Art. no. baw113.

[102] P.-T. Lai, Y.-Y. Lo, M.-S. Huang, Y.-C. Hsiao, and R. T.-H. Tsai, "BelSmile: A biomedical semantic role labeling approach for extracting biological expression language from text," *Database*, vol. 2016, Jan. 2016, Art. no. baw064.

[103] S. Madan, J. Szostak, R. K. Elayavilli, R. T.-H. Tsai, M. Ali, L. Qian, M. Rastegar-Mojarad, J. Hoeng, and J. Fluck, "The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track," *Database*, vol. 2019, Jan. 2019, Art. no. baz084.

[104] S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski, "BioInfer: A corpus for information extraction in the biomedical domain," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–24, Dec. 2007.

[105] P. Thompson, S. A. Iqbal, J. McNaught, and S. Ananiadou, "Construction of an annotated corpus to support biomedical information extraction," *BMC Bioinf.*, vol. 10, no. 1, p. 349, Dec. 2009.

[106] J.-D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa, "The genia event and protein coreference tasks of the BioNLP shared task 2011," *BMC Bioinf.*, vol. 13, no. S-11, p. S1, Jun. 2012.

[107] T. Ohta, S. Pyysalo, and J. Tsujii, "Overview of the epigenetics and post-translational modifications (EPI) task of BioNLP shared task 2011," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Jun. 2011, pp. 16–25.

[108] S. Pyysalo, T. Ohta, M. Miwa, H. C. Cho, J. Tsujii, and S. Ananiadou, "Event extraction across multiple levels of biological organization," *Bioinformatics*, vol. 28, no. 18, pp. 575–581, 2012.

[109] J. Kim, Y. Wang, and Y. Yamamoto, "The genia event extraction shared task, 2013 edition-overview," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 8–15.

[110] T. Ohta, S. Pyysalo, R. Rak, A. Rowley, H. Chun, S. Jung, S. Choi, S. Ananiadou, and J. Tsujii, "Overview of the pathway curation (PC) task of BioNLP shared task 2013," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 67–75.

[111] J. Kim, X. Han, V. Lee, and D. Rebholz-Schuhmann, "GRO task: Populating the gene regulation ontology with events and relations," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 50–57.

[112] R. Bossy, P. Bessières, and C. Nédellec, "BioNLP shared task 2013—An overview of the genic regulation network task," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 153–160.

[113] Y. Chen, "A transfer learning model with multi-source domains for biomedical event trigger extraction," *BMC Genomics*, vol. 22, no. 1, pp. 1–18, Dec. 2021.

[114] M. Miwa, P. Thompson, and S. Ananiadou, "Boosting automatic event extraction from the literature using domain adaptation and coreference resolution," *Bioinformatics*, vol. 28, no. 13, pp. 1759–1765, 2012.

[115] S. Zheng, W. Cao, W. Xu, and J. Bian, "Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 337–346.

[116] S. Zheng, W. Cao, W. Xu, and J. Bian, "Revisiting the evaluation of end-to-end event extraction," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 4609–4617.

[117] D. Zhou, L. Chen, and Y. He, "A simple Bayesian modelling approach to event extraction from Twitter," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, Baltimore, MD, USA, 2014, pp. 700–705.

[118] D. Zhou, L. Chen, and Y. He, "An unsupervised framework of exploring events on Twitter: Filtering, extraction and categorization," in *Proc. 29th AAAI Conf. Artif. Intell.*, Austin, TX, USA, B. Bonet and S. Koenig, Eds., Jan. 2015, pp. 2468–2475.

[119] D. Zhou, X. Zhang, and Y. He, "Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Valencia, Spain, M. Lapata, P. Blunsom, and A. Koller, Eds., Apr. 2017, pp. 808–817.

[120] R. Wang, D. Zhou, and Y. He, "Open event extraction from online text using a generative adversarial network," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 282–291.

[121] N. Chambers and D. Jurafsky, "Template-based information extraction without the templates," in *Proc. Conf. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Jun. 2011, pp. 976–986.

[122] K. Nguyen, X. Tannier, O. Ferret, and R. Besançon, "Generative event schema induction with entity disambiguation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process. (ACL)*, Beijing, China, Jul. 2015, pp. 188–197.

[123] Q. Yuan, X. Ren, W. He, C. Zhang, X. Geng, L. Huang, H. Ji, C. Lin, and J. Han, "Open-schema event profiling for massive news corpora," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Turin, Italy, A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, and H. Wang, Eds., Oct. 2018, pp. 587–596.

[124] X. Liu, H. Huang, and Y. Zhang, "Open domain event extraction using neural latent variable models," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul./Aug. 2019, pp. 2860–2871.

[125] M. Li, Q. Zeng, Y. Lin, K. Cho, H. Ji, J. May, N. Chambers, and C. R. Voss, "Connecting the dots: Event graph schema induction with path language modeling," in *Proc. Conf. Empirical Methods Natural Lang. (EMNLP)*, Nov. 2020, pp. 684–695.

[126] L. Huang, T. Cassidy, X. Feng, H. Ji, C. R. Voss, J. Han, and A. Sil, "Liberal event extraction and event schema induction," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, Aug. 2016, pp. 258–268.

[127] Y. Yang, T. Pierce, and J. G. Carbonell, "A study of retrospective and on-line event detection," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Melbourne, VIC, Australia, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds., Aug. 1998, pp. 28–36.

[128] N. Stokes and J. Carthy, "Combining semantic and syntactic document classifiers to improve first story detection," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New Orleans, LA, USA, W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, Eds., Sep. 2001, pp. 424–425.

[129] Z. Li, B. Wang, M. Li, and W. Ma, "A probabilistic model for retrospective news event detection," in *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Salvador, Brazil, R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, Eds., 2005, pp. 106–113.

[130] M. Naughton, N. Kushmerick, and J. Carthy, "Event extraction from heterogeneous news sources," in *Proc. AAAI Workshop Event Extraction Synth.*, 2006, pp. 1–6.

[131] S. Ribeiro, O. Ferret, and X. Tannier, "Unsupervised event clustering and aggregation from newswire and web articles," in *Proc. Workshop, Natural Lang. Process. Meets Journalism (NLPmJ@EMNLP)*, Copenhagen, Denmark, O. Popescu and C. Strapparava, Eds., Sep. 2017, pp. 62–67.

[132] S. Yu and B. Wu, "Exploiting structured news information to improve event detection via dual-level clustering," in *Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC)*, Guangzhou, China, Jun. 2018, pp. 873–880.

[133] X. Du and C. Cardie, "Event extraction by answering (almost) natural questions," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 671–683.

[134] Q. Li, H. Peng, J. Li, Y. Ning, L. Wang, P. S. Yu, and Z. Wang, "Reinforcement learning-based dialogue guided event extraction to exploit argument relations," Jun. 2021, *arXiv:2106.12384*.

[135] X. D. Wang, L. Weber, and U. Leser, "Biomedical event extraction as multi-turn question answering," in *Proc. 11th Int. Workshop Health Text Mining Inf. Anal. (LOUHI@EMNLP)*, E. Holderness, A. Jimeno-Yepes, A. Lavelli, A. Minard, J. Pustejovsky, and F. Rinaldi, Eds., Nov. 2020, pp. 88–96.

[136] G. Paolini, B. Athiwaratkun, J. Krone, J. Ma, A. Achille, R. Anubhai, C. N. dos Santos, B. Xiang, and S. Soatto, "Structured prediction as translation between augmented natural languages," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021.

[137] J. Zhang, Y. Qin, Y. Zhang, M. Liu, and D. Ji, "Extracting entities and events as a single task using a transition-based neural model," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, Macao, China, S. Kraus, Ed., Aug. 2019, pp. 5422–5428.

[138] K.-H. Huang and N. Peng, "Document-level event extraction with efficient end-to-end learning of cross-event dependencies," in *Proc. 3rd Workshop Narrative Understand.*, 2021, pp. 36–47.

[139] K. Huang, M. Yang, and N. Peng, "Biomedical event extraction on graph edge-conditioned attention networks with hierarchical knowledge graphs," in *Proc. Conf. Empirical Methods Natural Lang. Process., Findings (EMNLP)*, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 1277–1285.

[140] A. Ramponi, R. van der Goot, R. Lombardo, and B. Plank, "Biomedical event extraction as sequence labeling," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 5357–5367.

[141] J. Liu, Y. Chen, K. Liu, Y. Jia, and Z. Sheng, "How does context matter? On the robustness of event detection with context-selective mask generalization," in *Proc. Conf. Empirical Methods Natural Lang. Process., Findings (EMNLP)*, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 2523–2532.

[142] A. Subburathinam, D. Lu, H. Ji, J. May, S. Chang, A. Sil, and C. R. Voss, "Cross-lingual structure transfer for relation and event extraction," in *Proc. Conf. Empirical Methods Natural Lang., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, Nov. 2019, pp. 313–325.

[143] W. U. Ahmad, N. Peng, and K. Chang, "GATE: Graph attention transformer encoder for cross-lingual relation and event extraction," in *Proc. 35th AAAI Conf. Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell., 11th Symp. Educ. Adv. Artif. Intell. (EAAI)*, Feb. 2021, pp. 12462–12470.

[144] S. Wei, I. Korostil, J. Nothman, and B. Hachey, "English event detection with translated language features," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, BC, Canada, R. Barzilay and M. Kan, Eds., Jul./Aug. 2017, pp. 293–298.

[145] Z. Chen and H. Ji, "Can one language bootstrap the other: A case study on event extraction," in *Proc. NAACL HLT Workshop Semi-Supervised Learn. Natural Lang. Process.*, 2009, pp. 66–74.

[146] Z. Zhu, S. Li, G. Zhou, and R. Xia, "Bilingual event extraction: A case study on trigger type determination," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Baltimore, MD, USA, Jun. 2014, pp. 842–847.

[147] J. Liu, Y. Chen, K. Liu, and J. Zhao, "Neural cross-lingual event detection with minimal parallel resources," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 738–748.

[148] H. Yang, Y. Chen, K. Liu, Y. Xiao, and J. Zhao, "DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data," in *Proc. ACL, Syst. Demonstrations*, Melbourne, VIC, Australia, F. Liu and T. Solorio, Eds., Jul. 2018, pp. 50–55.

[149] S. Rao, D. Marcu, K. Knight, and H. Daumé, III, "Biomedical event extraction using abstract meaning representation," in *Proc. BioNLP*, Vancouver, BC, Canada, K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, Eds., Aug. 2017, pp. 126–135.

[150] Y. Zeng, Y. Feng, R. Ma, Z. Wang, R. Yan, C. Shi, and D. Zhao, "Scale up event extraction learning via automatic training data generation," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, S. A. McIlraith and K. Q. Weinberger, Eds., Feb. 2018, pp. 6045–6052.

[151] J. Araki and T. Mitamura, "Open-domain event detection using distant supervision," in *Proc. 27th Int. Conf. Comput. Linguistics (COLING)*, Santa Fe, NM, USA, E. M. Bender, L. Derczynski, and P. Isabelle, Eds., Aug. 2018, pp. 878–891.

[152] S. Liu, Y. Chen, S. He, K. Liu, and J. Zhao, "Leveraging FrameNet to improve automatic event detection," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, Aug. 2016, pp. 2134–2143.

[153] S. Liu, Y. Chen, K. Liu, and J. Zhao, "Exploiting argument information to improve event detection via supervised attention mechanisms," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, BC, Canada, R. Barzilay and M. Kan, Eds., Jul./Aug. 2017, pp. 1789–1798.

[154] T. H. Nguyen, L. Fu, K. Cho, and R. Grishman, "A two-stage approach for extending event detection to new types via neural networks," in *Proc. 1st Workshop Represent. Learn. NLP Rep4NLP@ACL*, Berlin, Germany, P. Blunsom, K. Cho, S. B. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, and S. W. Yih, Eds., Aug. 2016, pp. 158–165.

[155] H. Peng, Y. Song, and D. Roth, "Event detection and co-reference with minimal supervision," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Austin, TX, USA, J. Su, X. Carreras, and K. Duh, Eds., Nov. 2016, pp. 392–402.

[156] S. Deng, N. Zhang, J. Kang, Y. Zhang, W. Zhang, and H. Chen, "Meta-learning with dynamic-memory-based prototypical network for few-shot event detection," in *Proc. 13th ACM Int. Conf. Web Search Data Mining (WSDM)*, Houston, TX, USA, Feb. 2020, pp. 151–159.

[157] V. D. Lai, T. H. Nguyen, and F. Dernoncourt, "Extensively matching for few-shot learning event detection," in *Proc. NUSE@ACL*, 2020, pp. 38–45.

[158] S. Shen, T. Wu, G. Qi, Y. Li, G. Haffari, and S. Bi, "Adaptive knowledge-enhanced Bayesian meta-learning for few-shot event detection," in *Proc. Findings Assoc. Comput. Linguistics, ACL/IJCNLP*, Aug. 2021, pp. 2417–2429.

[159] L. Huang, H. Ji, K. Cho, I. Dagan, S. Riedel, and C. R. Voss, "Zero-shot transfer learning for event extraction," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, VIC, Australia, I. Gurevych and Y. Miyao, Eds., Jul. 2018, pp. 2160–2170.

[160] E. Boros, "Neural methods for event extraction. (Méthodes neuronales pour l'extraction d'événements)," Ph.D. dissertation, Univ. Paris-Saclay, Gif-sur-Yvette, France, 2018.

[161] S. Liao and R. Grishman, "Using prediction from sentential scope to build a pseudo co-testing learner for event extraction," in *Proc. 5th Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, Chiang Mai, Thailand, Nov. 2011, pp. 714–722.

[162] X. Han, J.-J. Kim, and C. K. Kwoh, "Active learning for ontological event extraction incorporating named entity recognition and unknown word handling," *J. Biomed. Semantics*, vol. 7, no. 1, p. 22, Dec. 2016.

[163] T. Munkhdalai, O.-E. Namsrai, and K. H. Ryu, "Self-training in significance space of support vectors for imbalanced biomedical event data," *BMC Bioinf.*, vol. 16, no. S-7, p. S6, 2015.

[164] X. Wang, X. Han, Z. Liu, M. Sun, and P. Li, "Adversarial training for weakly supervised event detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 998–1008.

[165] S. Liao and R. Grishman, "Can document selection help semi-supervised learning? A case study on event extraction," in *Proc. Conf. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 260–265.

[166] J. Ferguson, C. Lockard, D. S. Weld, and H. Hajishirzi, "Semi-supervised event extraction with paraphrase clusters," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, New Orleans, LA, USA, M. A. Walker, H. Ji, and A. Stent, Eds., Jun. 2018, pp. 359–364.

[167] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in *Proc. 11th Nat. Conf. Artif. Intell.*, Washington, DC, USA, R. Fikes and W. G. Lehnert, Eds., Jul. 1993, pp. 811–816.

[168] E. Riloff and J. Shoen, "Automatically acquiring conceptual patterns without an annotated corpus," in *Proc. 3rd Workshop Very Large Corpora (VLC@ACL)*, Cambridge, MA, USA, D. Yarowsky and K. Church, Eds., Jun. 1995, pp. 1–14.

[169] J.-T. Kim and D. I. Moldovan, "Acquisition of linguistic patterns for knowledge-based information extraction," *IEEE Trans. Knowl. Data Eng.*, vol. 7, no. 5, pp. 713–724, Oct. 1995.

[170] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen, "Automatic acquisition of domain knowledge for information extraction," in *Proc. 18th Int. Conf. Comput. Linguistics (COLING)*. Saarbrücken, Germany: Univ. Saarlandes, Jul./Aug. 2000, pp. 940–946.

[171] R. Yangarber, "Counter-training in discovery of semantic patterns," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguistics*, E. W. Hinrichs and D. Roth, Eds. Sapporo, Japan: Sapporo Convention Center, Jul. 2003, pp. 343–350.

[172] M. Stevenson and M. A. Greenwood, "A semantic approach to IE pattern induction," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, K. Knight, H. T. Ng, and K. Oflazer, Eds. Ann Arbor, MI, USA: Univ. Michigan, Jun. 2005, pp. 379–386.

[173] W. Phillips and E. Riloff, "Exploiting role-identifying nouns and expressions for information extraction," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2007, pp. 468–473.

[174] R. Huang and E. Riloff, "Bootstrapped training of event extraction classifiers," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Avignon, France, W. Daelemans, M. Lapata, and L. Màrquez, Eds., Apr. 2012, pp. 286–295.

[175] S. Patwardhan and E. Riloff, "Effective information extraction with semantic affinity patterns and relevant regions," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, Prague, Czech Republic, J. Eisner, Ed., Jun. 2007, pp. 717–727.

[176] Mausam, M. Schmitz, S. Soderland, R. Bart, and O. Etzioni, "Open language learning for information extraction," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn. (EMNLP-CoNLL)*, Jeju Island, South Korea, Jul. 2012, pp. 523–534.

[177] Q. Bui, D. Campos, E. M. van Mulligen, and J. A. Kors, "A fast rule-based approach for biomedical event extraction," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 104–108.

[178] Q. L. Minh, N. T. Son, and H. B. Quoc, "A pattern approach for biomedical event annotation," in *Proc. BioNLP@ACL Shared Task*, 2011, pp. 149–150.

[179] H. Kilicoglu and S. Bergler, "Syntactic dependency based heuristics for biological event extraction," in *Proc. BioNLP@HLT-NAACL Shared Task*, 2009, pp. 119–127.

[180] H. Ji, R. Grishman, Z. Chen, and P. Gupta, "Cross-document event extraction and tracking: Task, evaluation, techniques and challenges," in *Proc. Recent Adv. Natural Lang. Process. (RANLP)*, Borovets, Bulgaria, G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, Eds., Sep. 2009, pp. 166–172.

[181] L. Sha, J. Liu, C. Lin, S. Li, B. Chang, and Z. Sui, "RBPB: Regularization-based pattern balancing method for event extraction," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, Aug. 2016, pp. 1224–1234.

[182] T. Zhang, H. Ji, and A. Sil, "Joint entity and event extraction with generative adversarial imitation learning," *Data Intell.*, vol. 1, no. 2, pp. 99–120, 2019.

[183] P. Chen, H. Yang, K. Liu, R. Huang, Y. Chen, T. Wang, and J. Zhao, "Reconstructing event regions for event extraction via graph attention networks," in *Proc. 1st Conf. Asia–Pacific Chapter Assoc. Comput. Linguistics, 10th Int. Joint Conf. Natural Lang. Process. (AACL/IJCNLP)*, Suzhou, China, K. Wong, K. Knight, and H. Wu, Eds., Dec. 2020, pp. 811–820.

[184] R. Huang and E. Riloff, "Peeling back the layers: Detecting event role fillers in secondary contexts," in *Proc. Conf. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Jun. 2011, pp. 1137–1147.

[185] X. Du and C. Cardie, "Document-level event role filler extraction using multi-granularity contextualized encoding," in *Proc. Assoc. Comput. Linguistics (ACL)*, 2020, pp. 8010–8020.

[186] J. Zhang, Q. He, and Y. Zhang, "Syntax grounded graph convolutional network for joint entity and event extraction," *Neurocomputing*, vol. 422, pp. 118–128, Jan. 2021.

[187] X. Pham, M. Q. Le, and B. Ho, "A hybrid approach for biomedical event extraction," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 121–124.

[188] H. Liu, K. Verspoor, D. C. Comeau, A. MacKinlay, and W. J. Wilbur, "Generalizing an approximate subgraph matching-based system to extract events in molecular biology and cancer genetics," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 76–85.

[189] H. Liu, K. Verspoor, D. C. Comeau, A. D. MacKinlay, and W. J. Wilbur, "Optimizing graph-based patterns to extract biomedical events from the literature," *BMC Bioinf.*, vol. 16, no. S16, pp. 1–15, Dec. 2015.

[190] D. Venugopal, C. Chen, V. Gogate, and V. Ng, "Relieving the computational bottleneck: Joint inference for event extraction with high-dimensional features," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, A. Moschitti, B. Pang, and W. Daelemans, Eds., Oct. 2014, pp. 831–843.

[191] S. Abdulkadhar, B. Bhasuran, and J. Natarajan, "Multiscale Laplacian graph kernel combined with lexico-syntactic patterns for biomedical event extraction from literature," *Knowl. Inf. Syst.*, vol. 63, no. 1, pp. 143–173, Jan. 2021.

[192] X. Feng, L. Huang, D. Tang, H. Ji, B. Qin, and T. Liu, "A language-independent neural network for event detection," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Berlin, Germany, Aug. 2016, pp. 66–71.

[193] Y. Hong, W. Zhou, J. Zhang, Q. Zhu, and G. Zhou, "Self-regulation: Employing a generative adversarial network to improve event detection," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, VIC, Australia, I. Gurevych and Y. Miyao, Eds., Jul. 2018, pp. 515–526.

[194] J. Liu, Y. Chen, and K. Liu, "Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 6754–6761.

[195] H. Liu, R. K. Elayavilli, and K. Verspoor, "From graphs to events: A subgraph matching approach for information extraction from biomedical text," in *Proc. BioNLP@ACL Shared Task*, 2011, pp. 164–172.

[196] Y. Wang, J. Wang, H. Lin, X. Tang, S. Zhang, and L. Li, "Bidirectional long short-term memory with CRF for detecting biomedical event trigger in FastText semantic space," *BMC Bioinf.*, vol. 19, no. S20, pp. 59–66, Dec. 2018.

[197] H. Wei, A. Zhou, Y. Zhang, F. Chen, W. Qu, and M. Lu, "Biomedical event trigger extraction based on multi-layer residual BiLSTM and contextualized word representations," *Int. J. Mach. Learn. Cybern.*, pp. 1–13, Apr. 2021.

[198] X. He, Y. Ren, P. Tai, and H. Shi, "A two-stage biomedical event trigger detection method based on hybrid neural network and sentence embeddings," *IEEE Access*, vol. 9, pp. 81926–81935, 2021.

[199] W. Zhao, J. Zhang, J. Yang, T. He, H. Ma, and Z. Li, "A novel joint biomedical event extraction framework via two-level modeling of documents," *Inf. Sci.*, vol. 550, pp. 27–40, Mar. 2021.

[200] T. H. Nguyen and R. Grishman, "Graph convolutional networks with argument-aware pooling for event detection," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, S. A. McIlraith and K. Q. Weinberger, Eds., Feb. 2018, pp. 5900–5907.

[201] L. Sha, F. Qian, B. Chang, and Z. Sui, "Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, S. A. McIlraith and K. Q. Weinberger, Eds., Feb. 2018, pp. 5916–5923.

[202] W. Yu, M. Yi, X. Huang, X. Yi, and Q. Yuan, "Make it directly: Event extraction based on tree-LSTM and bi-GRU," *IEEE Access*, vol. 8, pp. 14344–14354, 2020.

[203] L. Li, J. Zheng, J. Wan, D. Huang, and X. Lin, "Biomedical event extraction via long short term memory networks along dynamic extended tree," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Shenzhen, China, Dec. 2016, pp. 739–742.

[204] X. Yu, W. Rong, J. Liu, D. Zhou, Y. Ouyang, and Z. Xiong, "LSTM-based end-to-end framework for biomedical event extraction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2029–2039, Nov. 2020.

[205] L. Li, J. Wan, J. Zheng, and J. Wang, "Biomedical event extraction based on GRU integrating attention mechanism," *BMC Bioinf.*, vol. 19, no. S9, pp. 177–184, Aug. 2018.

[206] S. Duan, R. He, and W. Zhao, "Exploiting document level information to improve event detection via recurrent neural networks," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, Taipei, Taiwan, G. Kondrak and T. Watanabe, Eds., Nov./Dec. 2017, pp. 352–361.

[207] T. M. Nguyen and T. H. Nguyen, "One for all: Neural joint modeling of entities and events," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 6851–6858.

[208] Y. Zhang, G. Xu, Y. Wang, X. Liang, L. Wang, and T. Huang, "Empower event detection with bi-directional neural language model," *Knowl.-Based Syst.*, vol. 167, pp. 87–97, Mar. 2019.

[209] Y. Zhao, X. Jin, Y. Wang, and X. Cheng, "Document embedding enhanced event detection with hierarchical and supervised attention," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, VIC, Australia, I. Gurevych and Y. Miyao, Eds., Jul. 2018, pp. 414–419.

[210] S. Liu, Y. Li, F. Zhang, T. Yang, and X. Zhou, "Event detection without triggers," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 735–744.

[211] Y. Chen, H. Yang, K. Liu, J. Zhao, and Y. Jia, "Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Oct./Nov. 2018, pp. 1267–1276.

[212] J. Björne and T. Salakoski, "Biomedical event extraction using convolutional neural networks and dependency parsing," in *Proc. BioNLP Workshop*, Melbourne, VIC, Australia, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds., Jul. 2018, pp. 98–108.

[213] A. Wang, J. Wang, H. Lin, J. Zhang, Z. Yang, and K. Xu, "A multiple distributed representation method based on neural network for biomedical event extraction," *BMC Med. Informat. Decis. Making*, vol. 17, no. S3, pp. 59–66, Dec. 2017.

[214] T. H. Nguyen and R. Grishman, "Event detection and domain adaptation with convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process., Asian Fed. Natural Lang. Process. (ACL)*, Beijing, China, Jul. 2015, pp. 365–371.

[215] T. H. Nguyen and R. Grishman, "Modeling skip-grams for event detection with convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Austin, TX, USA, J. Su, X. Carreras, and K. Duh, Eds., Nov. 2016, pp. 886–891.

[216] R. Huang and E. Riloff, "Modeling textual cohesion for event extraction," in *Proc. 26th AAAI Conf. Artif. Intell.*, Toronto, ON, Canada, J. Hoffmann and B. Selman, Eds., Jul. 2012, pp. 1–7.

[217] Q. Li, H. Ji, and L. Huang, "Joint event extraction via structured prediction with global features," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Sofia, Bulgaria, Aug. 2013, pp. 73–82.

[218] A. Judea and M. Strube, "Incremental global event extraction," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, Osaka, Japan, N. Calzolari, Y. Matsumoto, and R. Prasad, Eds., Dec. 2016, pp. 2279–2289.

[219] Q. Li, H. Ji, Y. Hong, and S. Li, "Constructing information networks using one single model," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, A. Moschitti, B. Pang, and W. Daelemans, Eds., Oct. 2014, pp. 1846–1851.

[220] B. Yang and T. M. Mitchell, "Joint extraction of events and entities within a document context," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL HLT)*, San Diego, CA, USA, K. Knight, A. Nenkova, and O. Rambow, Eds., Jun. 2016, pp. 289–299.

[221] S. Riedel, H. Chun, T. Takagi, and J. Tsujii, "A Markov logic approach to bio-molecular event extraction," in *Proc. BioNLP Workshop Companion Volume Shared Task (BioNLP@HLT-NAACL)*, Boulder, CO, USA, Jun. 2009, pp. 41–49.

[222] S. Riedel, R. Sætre, H.-W. Chun, T. Takagi, and J. Tsujii, "Bio-molecular event extraction with Markov logic," *Comput. Intell.*, vol. 27, no. 4, pp. 558–582, 2011.

[223] D. McClosky, M. Surdeanu, and C. D. Manning, "Event extraction as dependency parsing," in *Proc. Conf. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Jun. 2011, pp. 1626–1635.

[224] A. Vlachos and M. Craven, "Biomedical event extraction from abstracts and full papers using search-based structured prediction," *BMC Bioinf.*, vol. 13, no. S-11, p. S5, 2012.

[225] S. Riedel, D. McClosky, M. Surdeanu, A. McCallum, and C. D. Manning, "Model combination for event extraction in BioNLP 2011," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Jun. 2011, pp. 51–55.

[226] D. McClosky, S. Riedel, M. Surdeanu, A. McCallum, and C. D. Manning, "Combining joint models for biomedical event extraction," *BMC Bioinf.*, vol. 13, no. S-11, p. S9, Jun. 2012.

[227] L. Li, S. Liu, M. Qin, Y. Wang, and D. Huang, "Extracting biomedical event with dual decomposition integrating word embeddings," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 4, pp. 669–677, Jul./Aug. 2016.

[228] S. Liu, K. Liu, S. He, and J. Zhao, "A probabilistic soft logic based approach to exploiting latent and global information in event classification," in *Proc. 30th AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, D. Schuurmans and M. P. Wellman, Eds., Feb. 2016, pp. 2993–2999.

[229] S. Patwardhan and E. Riloff, "A unified model of phrasal and sentential evidence for information extraction," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, Aug. 2009, pp. 151–160.

[230] D. Ahn, "The stages of event extraction," in *Proc. Workshop Annotating Reasoning About Time Events (ARTE)*, 2006, pp. 1–8.

[231] H. L. Chieu and H. T. Ng, "A maximum entropy approach to information extraction from semi-structured and free text," in *Proc. 18th Nat. Conf. Artif. Intell., 14th Conf. Innov. Appl. Artif. Intell.*, Edmonton, AB, Canada, R. Dechter, M. J. Kearns, and R. S. Sutton, Eds., Jul./Aug. 2002, pp. 786–791.

[232] H. Ji and R. Grishman, "Refining event extraction through cross-document inference," in *Proc. 46th Annu. Meeting Assoc. Comput. Linguistics*, Columbus, OH, USA, K. R. McKeown, J. D. Moore, S. Teufel, J. Allan, and S. Furui, Eds., Jun. 2008, pp. 254–262.

[233] S. Liao and R. Grishman, "Using document level cross-event inference to improve event extraction," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Uppsala, Sweden, J. Hajic, S. Carberry, and S. Clark, Eds., Jul. 2010, pp. 789–797.

[234] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski, "Extracting complex biological events with rich graph-based feature sets," in *Proc. Workshop BioNLP Shared Task (BioNLP)*, 2009, pp. 10–18.

[235] J. Björne and T. Salakoski, "TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 16–25.

[236] M. Miwa, R. Sætre, J.-D. Kim, and J. Tsujii, "Event extraction with complex event classification using rich features," *J. Bioinf. Comput. Biol.*, vol. 8, no. 1, pp. 131–146, 2010.

[237] J. Björne and T. Salakoski, "TEES 2.2: Biomedical event extraction for diverse corpora," *BMC Bioinf.*, vol. 16, no. S16, pp. 1–20, Dec. 2015.

[238] J. Lever and S. J. Jones, "VERSE: Event and relation extraction in the BioNLP 2016 shared task," in *Proc. BioNLP Shared Task*, 2016, pp. 42–49.

[239] X. Liu, A. Bordes, and Y. Grandvalet, "Biomedical event extraction by multi-class classification of pairs of text entities," in *Proc. BioNLP@ACL Shared Task*, 2013, pp. 45–49.

[240] K. Hakala, S. V. Landeghem, T. Salakoski, Y. V. de Peer, and F. Ginter, "EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction," in *Proc. BioNLP Shared Task Workshop*, Sofia, Bulgaria, C. Nédellec, R. Bossy, J. Kim, J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, Eds., Aug. 2013, pp. 26–34.

[241] C. Aone and M. Ramos-Santacruz, "REES: A large-scale relation and event extraction system," in *Proc. 6th Appl. Natural Lang. Process. Conf. (ANLP)*, Seattle, WA, USA, Apr./May 2000, pp. 76–83.

[242] F. de Sá Mesquita, J. Schmidek, and D. Barbosa, "Effectiveness and efficiency of open relation extraction," in *Proc. EMNLP*, 2013, pp. 447–457.

[243] N. Bhutani, H. V. Jagadish, and D. R. Radev, "Nested propositions in open information extraction," in *Proc. EMNLP*, 2016, pp. 55–64.

[244] S. B. de Vroe, L. Guillou, M. Stanojević, N. McKenna, and M. Steedman, "Modality and negation in event extraction," in *Proc. 4th Workshop Challenges Appl. Automated Extraction Socio-Political Events From Text (CASE)*, 2021, pp. 1–12.

[245] K. B. Cohen, K. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner, Jr., E. White, H. Tipney, and L. Hunter, "High-precision biological event extraction with a concept recognizer," in *Proc. BioNLP Workshop Companion Volume Shared Task (BioNLP@HLT-NAACL)*, Boulder, CO, USA, Jun. 2009, pp. 50–58.

[246] A. Casillas, A. D. de Ilarraza, K. Gojenola, M. Oronoz, and G. Rigau, "Using kybots for extracting events in biomedical texts," in *Proc. BioNLP Shared Task Workshop*, Portland, OR, USA, J. Tsujii, J. Kim, and S. Pyysalo, Eds., Jun. 2011, pp. 138–142.

[247] D. McClosky, E. Charniak, and M. Johnson, "Automatic domain adaptation for parsing," in *Proc. Hum. Lang. Technol., Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, Jun. 2010, pp. 28–36.

[248] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "ATOMIC: An atlas of machine commonsense for if-then reasoning," in *Proc. 33rd AAAI Conf. Artif. Intell., 31th Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 3027–3035.

[249] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi, "COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs," in *Proc. 35th AAAI Conf. Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell., 11th Symp. Educ. Adv. Artif. Intell., (EAAI), Virtual Event*. AAAI Press, Feb. 2021, pp. 6384–6392. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16792 and https://dblp.org/rec/conf/aaai/HwangBBDSBC21.bib

[250] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "A survey on open information extraction," in *Proc. COLING*, 2018, pp. 3866–3878.

[251] D. Zhou and Y. He, "Biomedical events extraction using the hidden vector state model," *Artif. Intell. Med.*, vol. 53, no. 3, pp. 205–213, 2011.

[252] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[253] J. Björne, F. Ginter, and T. Salakoski, "University of Turku in the BioNLP'11 shared task," *BMC Bioinf.*, vol. 13, no. S-11, p. S4, 2012.

[254] M. Miwa and S. Ananiadou, "Adaptable, high recall, event extraction system with minimal configuration," *BMC Bioinf.*, vol. 16, no. S10, pp. 1–11, Dec. 2015.

[255] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski, "Complex event extraction at PubMed scale," *Bioinformatics*, vol. 26, no. 12, pp. 382–390, 2010.

[256] S. Van Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, and F. Ginter, "Large-scale event extraction from literature with multi-level gene normalization," *PLoS ONE*, vol. 8, no. 4, Apr. 2013, Art. no. e55814.

[257] N. T. Ngo, T. N. Nguyen, and T. H. Nguyen, "Learning to select important context words for event detection," in *Proc. 24th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, in Lecture Notes in Comput. Science, vol. 12085, Singapore, H. W. Lauw, R. C. Wong, A. Ntoulas, E. Lim, S. Ng, and S. J. Pan, Eds. Cham, Switzerland: Springer, May 2020, pp. 756–768.

[258] W. Zhao, Y. Zhang, X. Jiang, T. He, F. Liu, and N. Li, "Efficient multiple biomedical events extraction via reinforcement learning," *Bioinformatics*, vol. 37, no. 13, pp. 1891–1899, Jul. 2021.

[259] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. 47th Annu. Meeting Assoc. Comput. Linguistics, 4th Int. Joint Conf. Natural Lang. Process. (AFNLP)*, Singapore, K. Su, J. Su, and J. Wiebe, Eds., Aug. 2009, pp. 1003–1011.

[260] B. Roth, T. Barth, M. Wiegand, and D. Klakow, "A survey of noise reduction methods for distant supervision," in *Proc. Workshop Automated Knowl. Base Construct. AKBC@CIKM*, San Francisco, CA, USA, F. M. Suchanek, S. Riedel, S. Singh, and P. P. Talukdar, Eds., Oct. 2013, pp. 73–78.

[261] J. Aguilar, C. Beller, P. McNamee, B. V. Durme, S. M. Strassel, Z. Song, and J. Ellis, "A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards," in *Proc. 2nd Workshop EVENTS, Definition, Detection, Conf., Represent. (EVENTS@ACL)*, Baltimore, MD, USA, T. Mitamura, E. H. Hovy, and M. Palmer, Eds., Jun. 2014, pp. 45–53. [Online]. Available: https://arxiv.org/pdf/1910.10683.pdf

[262] X. Wang, Z. Wang, X. Han, W. Jiang, R. Han, Z. Liu, J. Li, P. Li, Y. Lin, and J. Zhou, "MAVEN: A massive general domain event detection dataset," in *Proc. EMNLP*, 2020, pp. 1652–1671.

[263] Y. Lin, C. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig, "Choosing transfer languages for cross-lingual learning," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul./Aug. 2019, pp. 3125–3135.

[264] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 30th Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Dec. 2017, pp. 5998–6008.

[265] Y. Goldberg, "Assessing BERT's syntactic abilities," Jan. 2019, *arXiv:1901.05287*.

[266] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. V. Durme, S. R. Bowman, D. Das, and E. Pavlick, "What do you learn from context? Probing for sentence structure in contextualized word representations," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, New Orleans, LA, USA, May 2019, pp. 1–17.

[267] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Proc. 32nd Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, H. M. Wallach, A. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., Dec. 2019, pp. 3261–3275.

[268] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 3613–3618.

[269] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[270] G. Miolo, G. Mantoan, and C. Orsenigo, "ELECTRAMed: A new pretrained language representation model for biomedical NLP," Apr. 2021, *arXiv:2104.09585*.

[271] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *CoRR*, vol. abs/2007.15779, Jul. 2020.

[272] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, and G. Altan-Bonnet, "SciFive: A text-to-text transformer model for biomedical literature," May 2021, *arXiv:2106.03598*.

[273] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds., Jun. 2019, pp. 4171–4186.

[274] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, 2010.

[275] H. Khachatrian, L. Nersisyan, K. Hambardzumyan, T. Galstyan, A. Hakobyan, A. Arakelyan, A. Rzhetsky, and A. Galstyan, "BioRelEx 1.0: Biological relation extraction benchmark," in *Proc. BioNLP@ACL*, 2019, pp. 176–190.

[276] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *J. Biomed. Inform.*, vol. 45, pp. 885–892, May 2012.

[277] P. Damodaran, "Parrot: Paraphrase generation for NLU," v1.0, 2021.

[278] E. Choi, J. Palomaki, M. Lamm, T. Kwiatkowski, D. Das, and M. Collins, "Decontextualization: Making sentences stand-alone," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 447–461, Apr. 2021.

[279] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–140:67, 2020.

[280] M. Bevilacqua, R. Blloshmi, and R. Navigli, "One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline," in *Proc. 35th AAAI Conf. Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell., 11th Symp. Educ. Adv. Artif. Intell. (EAAI)*, Feb. 2021, pp. 12564–12573.

[281] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., Jul. 2020, pp. 7871–7880.

[282] X. Xu, O. Dušek, V. Rieser, and I. Konstas, "AggGen: Ordering and aggregating while generating," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process. Long Papers*, vol. 1. Assoc. Comput. Linguistics, Aug. 2021, pp. 1419–1434. [Online]. Available: https://aclanthology.org/2021.acl-long.113, doi: 10.18653/v1/2021.acl-long.113.

[283] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," Jun. 2018, *arXiv:1806.08730*.

[284] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, "Zero-shot relation extraction via reading comprehension," in *Proc. 21st Conf. Comput. Natural Lang. Learn. (CoNLL)*, Vancouver, BC, Canada, R. Levy and L. Specia, Eds., Aug. 2017, pp. 333–342.

[285] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, "Style transfer through back-translation," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, VIC, Australia, I. Gurevych and Y. Miyao, Eds., Jul. 2018, pp. 866–876.

[286] P. Gupta, S. Rajaram, H. Schütze, and T. A. Runkler, "Neural relation extraction within and across sentence boundaries," in *Proc. 33rd AAAI Conf. Artif. Intell., 31st Innov. Appl. Artif. Intell. Conf., 9th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 6513–6520.

[287] Y. Luo, O. Uzuner, and P. Szolovits, "Bridging semantics and syntax with graph algorithms—State-of-the-art of extracting biomedical relations," *Briefings Bioinf.*, vol. 18, no. 1, pp. 160–178, 2017.

[288] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, A. Albahem, L. Cavedon, T. Cohn, T. Baldwin, and K. Verspoor, "ChEMU 2020: Natural language processing methods are effective for information extraction from chemical patents," *Frontiers Res. Metrics Anal.*, vol. 6, Mar. 2021, Art. no. 654438.

[289] J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen, and R. Barzilay, "Correction to automated chemical reaction extraction from scientific literature," *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 4124–4124, 2021, doi: 10.1021/acs.jcim.1c00834.

[290] S. Ananiadou, P. Thompson, R. Nawaz, J. McNaught, and D. B. Kell, "Event-based text mining for biology and functional genomics," *Briefings Funct. Genomics*, vol. 14, no. 3, pp. 213–230, May 2015.

[291] A. Spangher, N. Peng, J. May, and E. Ferrara, "Enabling low-resource transfer learning across COVID-19 corpora by combining event-extraction and co-training," in *Proc. 1st Workshop NLP COVID-ACL*, 2020, pp. 1–9.

[292] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the TAC 2010 knowledge base population track," in *Proc. 3rd Text Anal. Conf. (TAC)*, vol. 3, 2010, p. 3.

[293] K. Hakala, S. Van Landeghem, S. Kaewphan, T. Salakoski, Y. Van de Peer, and F. Ginter, "CyEVEX: Literature-scale network integration and visualization through cytoscape," *Proc. SMBM*, vol. 12, 2012, pp. 91–96.

[294] L. Hirschman, G. A. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey, T. Wiegers, and A. G. Winter, "Text mining for the biocuration workflow," *Database*, vol. 2012, pp. 1–10, Mar. 2012.

[295] M. Neves, A. Damaschun, N. Mah, F. Lekschas, S. Seltmann, H. Stachelscheid, J.-F. Fontaine, A. Kurtz, and U. Leser, "Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts," *Database*, vol. 2013, Apr. 2013, Art. no. bat020.

[296] M. Miwa, T. Ohta, R. Rak, A. Rowley, D. B. Kell, S. Pyysalo, and S. Ananiadou, "A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text," *Bioinformatics*, vol. 29, no. 13, pp. 44–52, 2013.

[297] Y. Tsuruoka, M. Miwa, K. Hamamoto, J. Tsujii, and S. Ananiadou, "Discovering and visualizing indirect associations between biomedical concepts," *Bioinformatics*, vol. 27, no. 13, pp. 111–119, 2011.

[298] S. Henry and B. T. McInnes, "Literature based discovery: Models, methods, and trends," *J. Biomed. Inform.*, vol. 74, pp. 20–32, Oct. 2017.

[299] T. Zhang, M. Chen, and A. A. T. Bui, "Diagnostic prediction with Sequence-of-sets representation learning for clinical events," in *Proc. 18th Int. Conf. Artif. Intell. Med. (AIME)*, in Lecture Notes in Computer Science, vol. 12299, Minneapolis, MN, USA, M. Michalowski and R. Moskovitch, Eds. Cham, Switzerland: Springer, Aug. 2020, pp. 348–358.

[300] J. Berant, V. Srikumar, P. Chen, A. V. Linden, B. Harding, B. Huang, P. Clark, and C. D. Manning, "Modeling biological processes for reading comprehension," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, A. Moschitti, B. Pang, and W. Daelemans, Eds., Oct. 2014, pp. 1499–1510.

[301] R. K. Srihari and W. Li, "A question answering system supported by information extraction," in *Proc. 6th Appl. Natural Lang. Process. Conf. (ANLP)*, Seattle, WA, USA, Apr./May 2000, pp. 166–172.

[302] E. Filatova and V. Hatzivassiloglou, "Event-based extractive summarization," in *Text Summarization Branches Out*. Barcelona, Spain: Assoc. Comput. Linguistics, Jul. 2004, pp. 104–111. [Online]. Available: https://aclanthology.org/W04-1017

[303] W. Li, L. He, and H. Zhuge, "Abstractive news summarization based on event semantic link network," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, Osaka, Japan, N. Calzolari, Y. Matsumoto, and R. Prasad, Eds., Dec. 2016, pp. 236–246.

[304] J. Zhang, K. Li, C. Yao, and Y. Sun, "Event-based summarization method for scientific literature," *Pers. Ubiquitous Comput.*, vol. 25, pp. 1–10, Apr. 2020.

[305] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "MS²: Multi-document summarization of medical studies," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Assoc. Comput. Linguistics, Nov. 2021, pp. 7494–7513. [Online]. Available: https://aclanthology.org/2021.emnlp-main.594

[306] Y. Miyao, T. Ohta, K. Masuda, Y. Tsuruoka, K. Yoshida, T. Ninomiya, and J. Tsujii, "Semantic retrieval for the accurate identification of relational concepts in massive textbases," in *Proc. 21st Int. Conf. Comput. Linguistics, 44th Annu. Meeting ACL (ACL)*, 2006, pp. 1017–1024.

[307] H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi, "Event2Mind: Commonsense inference on events, intents, and reactions," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Melbourne, VIC, Australia, I. Gurevych and Y. Miyao, Eds., Jul. 2018, pp. 463–473.

[308] S. Chaturvedi, H. Peng, and D. Roth, "Story comprehension for predicting what happens next," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Copenhagen, Denmark, M. Palmer, R. Hwa, and S. Riedel, Eds., Sep. 2017, pp. 1603–1614.

[309] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI)*, Buenos Aires, Argentina, Q. Yang and M. J. Wooldridge, Eds., Jul. 2015, pp. 2327–2333.

[310] M. Chen, H. Zhang, Q. Ning, M. Li, H. Ji, K. McKeown, and D. Roth, "Event-centric natural language processing," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process., Tutorial Abstr.* Assoc. Comput. Linguistics, Aug. 2021, pp. 6–14. [Online]. Available: https://aclanthology.org/2021.acl-tutorials.2, doi: 10.18653/v1/2021.acl-tutorials.2.

[311] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020, *arXiv:2009.06732*.

[312] O. Agarwal, H. Ge, S. Shakeri, and R. Al-Rfou, "Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Assoc. Comput. Linguistics, Jun. 2021, pp. 3554–3565. [Online]. Available: https://dblp.org/rec/conf/naacl/AgarwalGSA21.bib, doi: 10.18653/v1/2021.naacl-main.278.

[313] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," Feb. 2020, *arXiv:2002.08909*.

[314] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "Iterative refining of category profiles for nearest centroid cross-domain text classification," in *Proc. 6th Int. Joint Conf. Knowl. Discovery, Knowl. Eng., Knowl. Manage. (ICK)*, vol. 553, Rome, Italy, A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, and J. Filipe, Eds. Cham, Switzerland: Springer, Oct. 2014, pp. 50–67.

[315] G. Domeniconi, G. Moro, A. Pagliarani, and R. Pasolini, "On deep learning in cross-domain sentiment classification," in *Proc. 9th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage. (ICK)*, A. L. N. Fred and J. Filipe, Eds., vol. 1, 2017, pp. 50–60.

[316] G. Frisoni, G. Moro, and A. Carbonaro, "Learning interpretable and statistically significant knowledge from unlabeled corpora of social text messages: A novel methodology of descriptive text mining," in *Proc. 9th Int. Conf. Data Sci., Technol. Appl. (DATA)*, 2020, pp. 121–134.

[317] G. Frisoni and G. Moro, "Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge," in *Proc. DATA*, vol. 1446. Cham, Switzerland: Springer, 2020, pp. 293–318.

[318] G. Frisoni, G. Moro, and A. Carbonaro, "Unsupervised descriptive text mining for knowledge graph learning," in *Proc. 12th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage. (ICK)*, vol. 1, 2020, pp. 316–324.

[319] G. Frisoni, G. Moro, and A. Carbonaro, "Towards rare disease knowledge graph learning from social posts of patients," in *Proc. RIIFORUM*. Cham, Switzerland: Springer, 2020, pp. 577–589.

[320] G. Frisoni, G. Moro, G. Carlassare, and A. Carbonaro, "Unsupervised event graph representation and similarity learning on biomedical literature," *Sensors*, pp. 1–31, 2021.

[321] G. Moro and L. Valgimigli, "Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature," *Sensors*, vol. 21, no. 19, p. 6430, Sep. 2021.

[322] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Nov. 2020, pp. 5418–5426.

[323] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller, "Language models as knowledge bases?" in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 2463–2473.

[324] J. Davison, J. Feldman, and A. M. Rush, "Commonsense knowledge mining from pretrained models," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 1173–1178.

[325] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. 33th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 2020, pp. 1–75.

[326] B. He, D. Zhou, J. Xiao, X. jiang, Q. Liu, N. Jing Yuan, and T. Xu, "Integrating graph contextualized knowledge into pre-trained language models," 2019, *arXiv:1912.00147*.

[327] R. L. L. IV, N. F. Liu, M. E. Peters, M. Gardner, and S. Singh, "Barack's wife hillary: Using knowledge graphs for fact-aware language modeling," in *Proc. 57th Conf. Assoc. Comput. Linguistics (ACL)*, Florence, Italy, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., Jul./Aug. 2019, pp. 5962–5971.

[328] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[329] I. Baaj, J.-P. Poli, and W. Ouerdane, "Some insights towards a unified semantic representation of explanation for eXplainable artificial intelligence," in *Proc. 1st Workshop Interact. Natural Lang. Technol. Explainable Artif. Intell. (NLXAI)*, 2019, pp. 14–19.

[330] A. Fan, T. Lavril, E. Grave, A. Joulin, and S. Sukhbaatar, "Addressing some limitations of transformers with feedback memory," 2020, *arXiv:2002.09402*.

[331] S. Sukhbaatar, D. Ju, S. Poff, S. Roller, A. Szlam, J. Weston, and A. Fan, "Not all memories are created equal: Learning to forget by expiring," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, in Proceedings of Machine Learning Research, vol. 139, M. Meila and T. Zhang, Eds. PMLR, Jul. 2021, pp. 9902–9912. [Online]. Available: http://proceedings.mlr.press/v139/sukhbaatar21a.html and https://dblp.org/rec/conf/icml/SukhbaatarJPRSW21.bib

[332] P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NIPS*, 2020, pp. 1–19.

[333] G. Moro, A. Pagliarani, R. Pasolini, and C. Sartori, "Cross-domain & in-domain sentiment analysis with memory-based deep neural networks," in *Proc. 10th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. Knowl. Manage. (ICK)*, vol. 1, 2018, pp. 127–138.

[334] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf.idf," in *Proc. DATA*, vol. 584. Cham, Switzerland: Springer, 2015, pp. 39–58.

[335] G. Domeniconi, G. Moro, A. Pagliarani, K. Pasini, and R. Pasolini, "Job recommendation from semantic similarity of linkedin users' skills," in *Proc. 5th Int. Conf. Pattern Recognit. Appl. Methods (ICPRAM)*, 2016, pp. 270–277.

GIACOMO FRISONI received the B.S. and M.S. degrees in computer science and engineering from the University of Bologna, Italy, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering.

In 2020, he was among the worldwide selected program attendees at the Cornell, Maryland, Max Planck Pre-doctoral Research School. His research interests include text mining, natural language understanding, and knowledge graphs, focusing on medical applications.

Dr. Frisoni was a recipient of the International Conference on Data Science, Technology and Applications Best Paper Award in 2020. Recently, he received the con.Scienze Award for having written one of the ten Italian best scientific research works during the master's thesis.
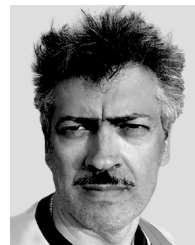
GIANLUCA MORO received the M.S. degree in computer science from the University of Bologna, Italy, in 1994, and the Ph.D. degree in computer engineering from the Department of Electronics, Computer Science and Systems, Italy, in 1999.

He is currently a Senior Research Associate at the Department of Computer Science and Engineering, University of Bologna, where he is also a Professor of text mining, data mining, and big data analytics. He is the Head of the Research Unit in text mining and natural language processing at Cesena Campus. He co-organized several editions of workshops on P2P and distributed systems at VLDB and AAMAS, edited five international books, and published more than 90 articles, some awarded as best papers in data science and information retrieval conferences. He has leaded national and international projects on data mining, text mining, and machine learning research topics and collaborates with public and private research organizations, among which the IBM Thomas J. Watson Research Center, USA.

ANTONELLA CARBONARO received the M.S. degree in computer science from the University of Bologna, Italy, in 1992, and the Ph.D. degree in intelligent systems from the Department of Information Engineering, University of Ancona, Italy, in 1997.

From 1997 to 1999, she won postdoctoral research grants on artificial intelligence topics. Since 2000, she has been first a Researcher and then an Associate Professor. She is currently a Confirmed Associate Professor at the Department of Computer Science and Engineering, University of Bologna. Her research interests include customization and content-based processing of information, data, and knowledge conceptualization with semantic web technologies and knowledge graph learning. Since 2015, she has been a member of the Technical-Scientific Committee of the Problem Solving Olympics at the Directorate-General for School Systems of the Ministry of Education, University and Research. She is also a member of the working group on Computer Science and School, CINI, the Italian Inter-University Consortium for Computer Science.

• • •