Joint Biomedical Event Extraction and Entity Linking via Iterative Collaborative Training

Xiaochu Li

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science Application

Lifu Huang, Chair
Chandan K. Reddy
Liqing Zhang

May 11, 2023
Blacksburg, Virginia

Keywords: Information extraction, Event extraction, Entity linking

Joint Biomedical Event Extraction and Entity Linking via Iterative Collaborative Training

Xiaochu Li

ABSTRACT

Biomedical entity linking and event extraction are two crucial tasks to support text understanding and retrieval in the biomedical domain. These two tasks intrinsically benefit each other: entity linking disambiguates the biomedical concepts by referring to external knowledge bases and the domain knowledge further provides additional clues to understand and extract the biological processes, while event extraction identifies a key trigger and entities involved to describe each biological process which also captures the structural context to better disambiguate the biomedical entities. However, previous research typically solves these two tasks separately or in a pipeline, leading to error propagation. What's more, it's even more challenging to solve these two tasks together as there is no existing dataset that contains annotations for both tasks. To solve these challenges, we propose joint biomedical entity linking and event extraction by regarding the event structures and entity references in knowledge bases as latent variables and updating the two task-specific models in an iterative training framework: (1) predicting the missing variables for each partially annotated dataset based on the current two task-specific models, and (2) updating the parameters of each model on the corresponding pseudo completed dataset. Experimental results on two benchmark datasets: Genia 2011 for event extraction and BC4GO for entity linking, show that our joint framework significantly improves the model for each individual task and outperforms the strong baselines for both tasks. We will make the code and model checkpoints publicly available once the paper is accepted.

Joint Biomedical Event Extraction and Entity Linking via Iterative Collaborative Training

Xiaochu Li

GENERAL AUDIENCE ABSTRACT

Biomedical entity linking and event extraction are essential tasks in understanding and retrieving information from biomedical texts. These tasks mutually benefit each other, as entity linking helps disambiguate biomedical concepts by leveraging external knowledge bases, while domain knowledge provides valuable insights for understanding and extracting biological processes. Event extraction, on the other hand, identifies triggers and entities involved in describing biological processes, capturing their contextual relationships for improved entity disambiguation. However, existing approaches often address these tasks separately or in a sequential manner, leading to error propagation. Furthermore, the joint solution becomes even more challenging due to the lack of datasets with annotations for both tasks.

To overcome these challenges, we propose a novel approach for jointly performing biomedical entity linking and event extraction. Our method treats the event structures and entity references in knowledge bases as latent variables and employs an iterative training framework. This framework involves predicting missing variables in partially annotated datasets based on the current task-specific models and updating the model parameters using the completed datasets. Experimental results on benchmark datasets, namely Genia 2011 for event extraction and BC4GO for entity linking, demonstrate the effectiveness of our joint framework. It significantly improves the performance of each individual task and outperforms strong baselines for both tasks.

This work is dedicated to my family, friends and Virginia Tech.

Acknowledgments

I am deeply grateful to my M.S. advisor, Dr. Lifu Huang, for his invaluable guidance and unwavering support throughout my research journey. His expertise, mentorship, and sincere encouragement have been instrumental in shaping the success of my thesis. I am deeply appreciative of the contributions and help from lab members, Minqian Liu and Zhiyang Xu. Their valuable suggestions and unwavering support have significantly enriched my work.

I extend my heartfelt thanks to my committee members, Dr. Liqing Zhang and Dr. Chandan K. Reddy, for their thoughtful feedback and constructive comments, which have played a pivotal role in improving the quality and depth of my thesis. Their expertise and insights have been invaluable in shaping the direction of this project.

Lastly, I would like to give my profound appreciation to my wife, daughter, and parents, for their constant encouragement and unwavering support throughout the years.

Contents

List of Figures

List of Tables

List of Abbreviations

EE Event Extraction

EL Entity Linking

GO Gene Ontology

NLP Natural Language Processing

Chapter 1

Introduction

1.1 Motivation

In today's society, we are living in an era that is occupied by an abundance of information. There are numerous sources that generate a constant influx of text data, including social networks, blogs, online news, search engines and etc. As a result, there has been a significant surge in research aimed at utilizing this immense and diverse collection of data to facilitate decision-making processes by disambiguating entities and detecting events. Furthermore, recent advancements in large-scale data processing technologies, encompassing both computational power and algorithms, have opened up innovative ways to leverage the vast amount of data. However, in order to employ the data for downstream applications, it is essential to convert the unstructured text into a structured format. This can be facilitated by utilizing natural language processing (NLP) technology, which automatically identifies instances of user-specified entities, relations, and events from unstructured text.

Biomedical NLP is a rapidly growing field that aims to extract relevant information from biomedical texts, such as scientific articles, electronic health records, and drug databases [1, 2, 3, 4]. The ability to accurately and efficiently identify entities and events in biomedical texts has numerous applications, such as biomedical research, drug discovery, clinical decision-making, and disease diagnosis [1, 2, 3, 4].

The field of biomedical research aims to discover new knowledge that can be applied to clinical uses such as disease diagnosis, prevention, and treatment. Traditionally, research has focused on studying individual biomedical entities, such as diseases, drugs, genes, proteins, and pathways. However, with the shift towards studying entire biological systems, the entity's functions, entity's roles, and relations between entities become even more important. For instance, systems biology explores the interplay among these entities and investigates their influence on the overall functionality and behavior of the system as a whole [5, 6, 7]. In biomedical networks, such as metabolic networks, networks regulating gene activity, and networks representing interactions between proteins, their interacting relationships are represented by genes, proteins, small molecules, and other entities [8, 9, 10]. Through computational analysis, valuable insights can be extracted from the information encapsulated within these networks, unveiling novel relationships, formulating hypotheses that can be empirically tested, and offering fresh perspectives on biological systems [8, 9, 10]. Network-based studies are also increasingly important in drug discovery, enabling researchers to better understand the relationships between drug-drug interaction and disease-associated genes [11, 12, 13].

The achievements of system biology, network biology, and drug discovery in the field of biomedicine heavily rely on the accessibility of precise, inclusive, and machine-interpretable information. Therefore, the acquisition and integration of knowledge have emerged as critical aspects of biomedical research. The number of biomedical research publications and the corresponding knowledge base have been experiencing exponential growth. To illustrate, the PubMed database currently encompasses over 35 million records and abstracts of biomedical literature [14]. Scientific literature plays a pivotal role as a fundamental knowledge source for biomedical research and drug discovery. However, with the rapid expansion of published biomedical scientific work and research, there is a growing concern that significant knowledge

and terms linking various biomedical entities may be inadvertently overlooked.

Keyword-based information retrieval methods are employed by popular search engines, such as Google and PubMed, which return all the papers and works that contain the searched words instead of comprehending the connections between concepts. This results in users being inundated with information that is often extraneous to their query, leading to time and energy being expended on filtering out irrelevant content. Furthermore, keyword-based search engines are incapable of providing responses to straightforward biomedical inquiries, such as: "what's the function of PLK1 (Polo-like Kinase 1) gene on cell cycle [15], what's the effect of drug STLC (S-trityl-L-cysteine) on mitosis in RPE1 cell [16]." To respond to such queries, it is necessary to extract machine-understandable knowledge from the literature [17, 18].

The acquisition of structured knowledge from literature necessitates the adoption of innovative methodologies. Biocuration is the process of converting information embedded in natural language within scientific reports into machine-understandable knowledge, a crucial undertaking in biological discovery and biomedical research. This task entails extensive manual curation efforts, where human curators diligently extract knowledge from the vast expanse of literature. One example is the Online Mendelian Inheritance in Man (OMIM), which is a knowledge base of gene-disease associations, containing information on approximately 16,000 genes and diseases [19, 20]. Nevertheless, the constantly evolving nature of biomedical research and the extensive reservoir of existing knowledge pose challenges to achieving complete knowledge bases. While the Unified Medical Language System (UMLS) serves as a prominent manually curated biomedical terminology source, its coverage remains fragmented [21]. Despite involving a large number of curators, each curation project has limited coverage and may not be up-to-date. The manual extraction of biomedical information from literature and converting it into machine-readable knowledge is challenging due to

the vast and complex nature of biomedical terminologies and knowledge [22, 23]. Furthermore, human curators are prone to errors and subjective bias [24, 25]. As a result, manually curated terminology and knowledge bases are considered incomplete.

Among the many tasks in biomedical NLP, entity linking and event extraction are particularly important as they are essential to aid domain experts in retrieving and organizing critical information related to gene functions, bio-molecule relations, and bio-molecule behaviors from the vast amount of unstructured texts [26, 27, 28]. Entity linking (EL) is considered to be the primary and vital step in the process of extracting structured information from unstructured text. Biomedical entity linking aims to automate this process by mapping biomedical mentions, such as diseases and drugs, to standardized medical entities in a knowledge base such as UMLS [29]. Event Extraction (EE) is a task that aims at extracting relational knowledge about given entity mentions occurring in an unstructured text. Informally, a biomedical event can be described as a structured representation of a biomedical process or occurrence that involves biomedical entities, such as proteins and genes [30]. In addition to identifying the interacting biomedical mentions, events seek to characterize the specific types of interactions taking place and the roles played by each mention in those interactions.

In this thesis, we focus on the problem of both biomedical entity linking and event extraction and propose an iterative training method to improve their performances. We first begin by providing an overview of biomedical entity linking and event extraction. Following this, we will outline the limitations and deficiencies of current state-of-the-art systems for every single task. Moving forward, we will propose a framework to iteratively improve both biomedical entity linking and event extraction, where each task-specific model incorporates the additional knowledge.

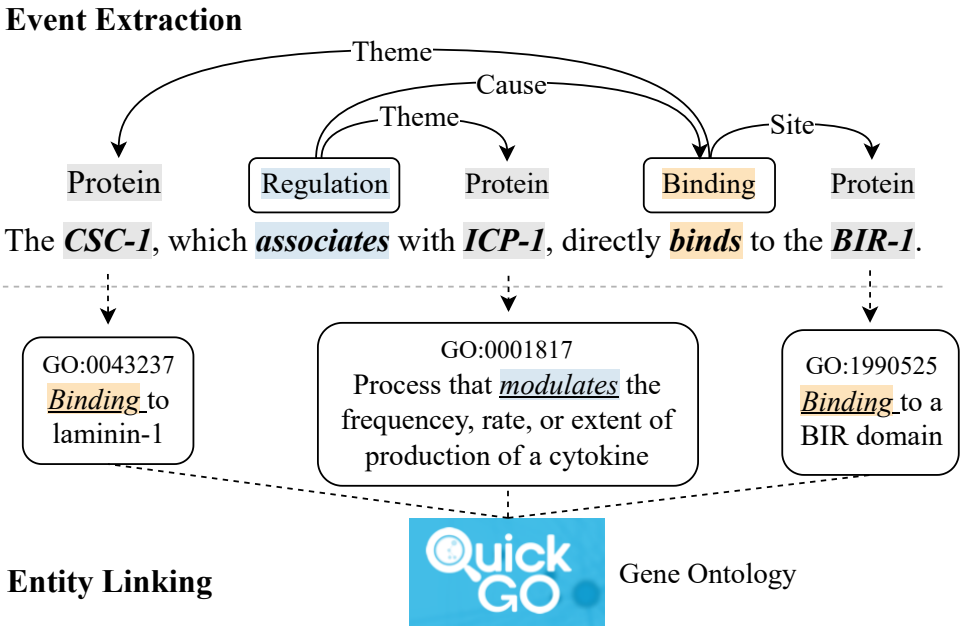

Figure 1.1: Illustration of biomedical entity linking (lower half) and event extraction (upper half) tasks given the same input. Below the input text, we show the definitions of each entity retrieved from Gene Ontology (GO) after running our entity linking model. We show the event types (in rounded boxes), entity types (without rounded boxes), and argument roles above the text. We highlight the event *Regulation* and its mention in blue, and the event *Binding* and its mention in orange. We also highlight the keywords in GO that are closely related to event extraction in corresponding colors.

1.2 Biomedical Entity Linking Overview

Biomedical entity linking (a.k.a. named-entity disambiguation) [31, 32, 33] aims to assign an entity mention in the text, such as genes, proteins, diseases and drugs, with a biomedical concept or term in reference biomedical knowledge bases, like Gene Ontology (GO) [34, 35], Unified Medical Language System (UMLS) [36], Universal Protein Resource (UniProt) [37], and the EMBL nucleotide sequence database [38]. However, this task poses a particular challenge as the same biomedical entity can have various names, including synonyms, morphological variations, and words with different orderings [39, 40]. Figure 1.1 (lower half) shows an example of biomedical entity linking, proteins in the given text can be linked to the

Gene Ontology (GO) knowledge base with a specific GO term which will provide biomedical definitions/functions to the linked entities. Accurate entity linking is crucial in the healthcare and biomedical research field to understand the biomedical context, as different biomedical concepts can be mentioned in similar ways. Failure to resolve these conflicts can result in a wrong perception of the entire context, leading to higher risks in biomedical decision-making. In addition, other applications that require automatic text indexing, such as medical information retrieval, predictive analysis, and question answering, can benefit from biomedical entity linking. Despite the potential benefits, entity linking in biomedical literature poses several challenges, such as misleading mentions, similar candidate entities, longer mentions and context, and domain-specific terms and abbreviations.

1.3 Biomedical Event Extraction Overview

Biomedical event extraction is the task of extracting relation knowledge given entity mentions in an unstructured text. Biomedical events are formal descriptions of biomedical processes or occurrences involving entities such as proteins. Events not only identify which biomedical mentions are interacting but also describe the type of interactions that occur and the role each mention plays [30]. This allows events to capture detailed descriptions of processes from complex biomedical statements found in scientific literature. Unlike entity linking, event extraction is semantically rich, The representation of the event is established based on *event triggers* in Figure 1.1. These triggers are tokens which are usually verbs or nominalized verbs that show the occurrence of a biomedical process of a particular type. Furthermore, events have numerous *arguments*, which are entities or other event triggers that take part in the events with a semantic function. For instance, figure 1.1 (top half) shows an example of biomedical event. Biomedical Event Extraction has numerous applications in biomedical

research, including drug discovery, systems biology, and personalized medicine. However, the complexity and diversity of biomedical knowledge make it challenging to extract relevant information accurately and efficiently.

1.4 Iterative Collaborative Training on Biomedical Entity Linking and Event Extraction

Despite the recent progress achieved in biomedical entity linking and event extraction, there are still several problems that remained unaddressed. In biomedical entity linking, the entity mentions can be highly ambiguous as one mention can be mapped to multiple distinct biomedical concepts, requiring the model to have a good understanding of the context of the mention. For example, *CSC-1* in Figure 1.1 can refer to a centromeric protein or a DNA [41].

Meanwhile, biomedical events usually have complex and nested structures and sufficient domain knowledge is required to capture biological processes and their participants. While each task has its own challenges, we find that these two tasks can be beneficial to each other: entity linking maps the mentions in the text to biomedical concepts in external knowledge bases and provides additional domain knowledge and semantic information (e.g., the definitions in Gene Ontology) for extracting biological processes, while event extraction identifies the key trigger and its associated arguments that can provide more structural context to narrow down the pool of candidates and better link the entities to the biomedical concepts in knowledge bases. As shown in Figure 1.1, the GO definition of the protein *CSC-1* clearly indicates the function of this protein is related to the biological process *binding*, which can help the event extraction model to infer the relationship between *CSC-1* and the *binding*

event. On the other hand, given that *CSC-1* is a *Theme* of *binding*, the entity linking can leverage such structural and precise context to better disambiguate the biological concept *CSC-1*.

Iterative training is an iterative process that involves training a machine learning algorithm on a subset of data and then using the trained model to make predictions on the remaining data. The predictions are then added to the training data, and the model is retrained on the expanded dataset. This process continues until the model reaches convergence, meaning that the predictions are no longer changing significantly.

Collaborative training, on the other hand, is a training approach that involves training a model on a combaination of labeled data and external knowledge sources, such as ontologies, knowledge graphs or other structured data. The model is trained on a specific dataset and learns to extract information based on the features present in that dataset. However, the model may not be able to capture all the nuances of the domain due to limited training data or other factors. Collaborative training can help to overcome this limitation by incorporating external knowledge sources or other models to provide additional information or insights. For example, in biomedical event extraction, the model can be trained on a dataset of annotated sentences, but may struggle to identify rare events or those that occur in unusual contexts. Collaborative training can help to address these issues by incorporating external knowledge sources such as ontologies, databases [42, 43, 44, 45], or other models trained on different datasets [46, 47]. The external sources can provide additional information about event types, entity relationships, or other relevant features that can improve the performance of the model.

Overall, iterative collaborative training can be a powerful technique in biomedical NLP that can be used to extract meaningful information from large amounts of unstructured text data. This technique has the potential to improve patient outcomes, leads to new scientific

discoveries, and advances our understanding of biomedical information.

1.5 Goals and contributions

While biomedical entity linking and event extraction intrinsically benefit each other, most existing works in biomedical information extraction ignore the close relationship between the two tasks and tackle them separately or in a pipeline, leading to the error propagation issue.

Besides, there is no existing dataset that contains annotations for both tasks. For example, the BC4GO dataset [35] only contains the annotations for entity linking, whereas the Genia 11 dataset [48] only has the annotations for event extraction. This makes it even more difficult to solve these two tasks together.

To address these challenges, we propose an iterative collaborative biomedical entity linking and event extraction framework, where each task-specific model incorporates the additional knowledge, i.e., the output from another model, to better perform task-specific prediction.

To iteratively improve the models specific to each task, we model the entity references in knowledge bases and event structures as latent variables and devise a collaborative training strategy that consists of two steps: (1) **Exstimation step** estimating the missing variables for each partially annotated dataset (e.g., event triggers and their argument roles in the entity linking dataset) using the current two task-specific models, and; (2) **Updation step** updating the parameters of each model on the corresponding dataset that has been augmented by the pseudo labels in the complementary task. These two steps are iteratively repeated until convergence, i.e., no further improvement in performance on the development set.

We extensively evaluate our approach on a biomedical entity linking dataset (i.e., BC4GO),

and an event extraction dataset (i.e., Genia 11). The experimental results and case study validate the effectiveness of our approach. Our main contributions in this work are summarized as follows:

- We propose an iterative biomedical entity linking and event extraction framework, namely JOINT4E-EL and JOINT4E-EE, where the two models can mutually improve each other.
- We design a collaborative training strategy to iteratively optimize two task-specific models such that each model can learn to leverage the information introduced by the other.
- Our joint framework significantly boosts the performance of the model for each individual task and outperforms previous strong baselines on both tasks.

Chapter 2

Related work

2.1 Related works in biomedical entity linking

Biomedical Entity Linking. Most recent state-of-the-art methods for biomedical entity linking are based on pre-trained BERT and consist of two steps: (1) candidate retrieval, which retrieves a small set of candidate references from a particular knowledge base; and (2) mention disambiguation and candidate ranking, which resolves the ambiguity of the mention based on the local context and refines the likelihood of each candidate reference with the fine-grained matching between mention and candidate [33, 49, 50]. These methods are not efficient enough as it requires two pipelined models (a retrieval and a ranking model) and have shown to not be able to generalize well on rare entities [33, 49]. Some recent studies have demonstrated that incorporating external information from biomedical knowledge bases, such as the latent type or semantic type information about mentions [31, 51], or infusing the domain-specific knowledge into the encoders with knowledge-aware pre-training tasks and objectives [52] can help improve the model performance on biomedical entity linking task [50]. While these studies mainly leverage the knowledge from external knowledge bases to improve biomedical entity linking, related tasks such as biomedical event extraction can also provide meaningful clues to disambiguate the meaning of the mentions in the local context, however, it has not been previously studied, especially in the biomedical domain.

2.2 Related work in biomedical event extraction

Biomedical Event Extraction Current approaches for biomedical event extraction mainly focus on extracting triggers and arguments in a pipeline [53, 54, 55, 56, 57]. Some studies also explore state-of-the-art neural methods with multiple classification layers to identify triggers, event types, arguments, and argument roles, respectively [42, 43, 53, 55]. Recently, [56] propose a sequence labeling framework by converting the extraction of event structures into a sequence labeling task by taking advantage of a multi-label aware encoding strategy. In addition, to improve the generalizability of event extraction, [57] establish a multi-turn question answering framework for event extraction by iteratively predicting answers for the template-based questions designed for event triggers, event arguments, and nested events. Several recent studies have also proposed to leverage external knowledge bases to disambiguate the biomedical terms in the local context and incorporate the knowledge, such as the definition or properties of the terms, into the event extraction process. Despite the success of these methods, they still suffer from error propagation in the pipeline frameworks, e.g., linking errors of biomedical terms in the local context will inform incorrect clues to the event extraction model and lead to a negative effect on the event predictions. Compared with all these studies, our JOINT4E-EE framework iteratively improves both biomedical entity linking and event extraction by leveraging the outputs from each other as additional input features.

Chapter 3

Iterative Biomedical event extraction and entity linking

3.1 Problem formulation

Biomedical Entity Linking. Given a text $\mathbf{x}^L = [x_1^L, x_2^L, ..., x_n^L]$ and a set of spans for all the entity mentions $\mathcal{M} = \{m_1, m_2, ..., m_p\}$ in $\mathbf{x}^L$, where n indicates the number of tokens and p indicates the number of mentions, biomedical entity linking maps each entity mention m_i to a particular entity concept $\hat{c}_i$ from a biomedical knowledge base. Taking the sentence in Figure 1.1 as an example, for each entity mention, such as *CSC-1*, a biomedical entity linking model will link it to a reference entity such as *GO:0043237* in the external knowledge base of *Gene Ontology*. Each entity in the knowledge base is represented with a unique GO ID and definition which is annotated by experts and Gene Ontology annotation tools[58, 59].

Biomedical Event Extraction. Biomedical event extraction consists of two subtasks: event detection and argument extraction. Given the input text $\mathbf{x}^E = [x_1^E, x_2^E, ..., x_n^E]$, the goal of *event detection* is to assign each token x_i^E in $\mathbf{x}^E$ with an event type τ_i that indicates a biological process in a predefined set of event types $\mathcal{T}$ or label it as *Other* if the token is not an event trigger. For each identified event trigger, *argument extraction* needs to assign each entity mention m_i in $\mathcal{M}$ with an argument role α_j or *Other* that indicates how the entity

participates in the biological process τ_i, where α_j belongs to a predefined set of argument role types $\mathcal{A}$. A mention is labeled as *Other* if it does not participate in the particular biological processes triggered by τ_i. As shown in Figure 1.1, given the sentence as input, biomedical event extraction aims to detect all the candidate triggers and their types, such as *associates* as a *Regulation* event mention and *binds* as a *Binding* event mention, and extract the arguments with arguments roles for each trigger, e.g., *ICP-1* is the *Theme* of the *associates* event while *BIR-1* is the *Site* of the *binds* event. Note that, each event mention can also be an argument in another event, for example, *associates* event is the *Cause* of the *binds* event. Thus, given a particular event trigger, we also predict an argument role α_j or *Other* for each of the other triggers.

3.2 Approach

In this section, we present our joint event extraction and entity linking framework that consists of (1) an entity-aware event extraction module, named JOINT4E-EE, that leverages the additional knowledge from knowledge bases, such as GO, UMLS, UniProt, and the EMBL nucleotide sequence database [34, 35, 36, 37, 38], to disambiguate the meaning of the biological terms in the input sentence so as to benefit the learning of the context and event extraction structures; and (2) an event-aware entity linking module, named JOINT4E-EL, which utilizes event structures to characterize the biological processes that each entity mention is involved and disambiguate its meaningful, so that we can better link each entity mention to the correct reference entity in the knowledge base. Since both JOINT4E-EL and JOINT4E-EE requires the output from the other task as additional input while there is no existing benchmark dataset containing annotations for both tasks, we further design a joint training framework to iteratively estimate the missing variables (i.e., event structures or

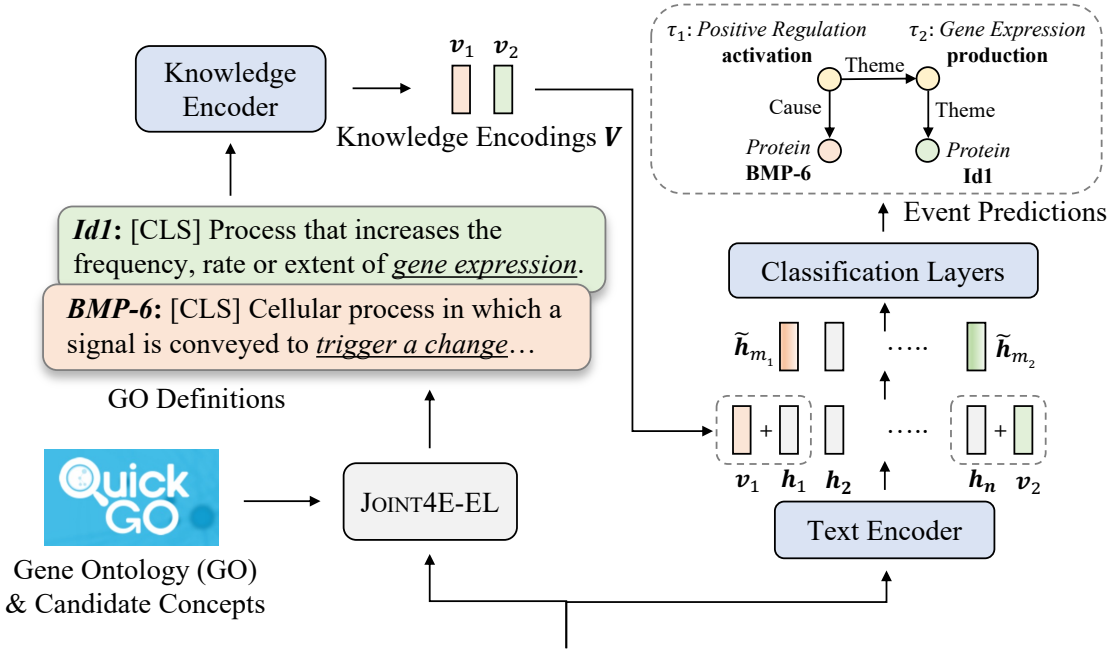

Figure 3.1: Illustration for our JOINT4E-EE for biomedical event extraction. JOINT4E-EE leverages the encoded GO definitions for each entity from the entity linking model JOINT4E-EL such that it has more domain knowledge to extract biological processes such as *Gene Expression* and its participant *Id1*.

entity references from external knowledge base) and optimize both JOINT4E-EL and JOINT4E-EE simultaneously. In the following, we first introduce the details of JOINT4E-EL and JOINT4E-EE in Section 3.2.2 and 3.2.1, and then elaborate on how we iteratively improve both task-specific models via an iterative learning schema in Section 3.2.3.

3.2.1 Entity-aware biomedical event extraction

Base Event Extraction Model. The base event extraction model takes a text $\mathbf{x}^E = [x_1^E, x_2^E, ..., x_n^E]$ and the set of all entity mentions $\mathcal{M}$ in $\mathbf{x}^E$ as inputs. We first encode $\mathbf{x}^E$ with a PLM encoder [60, 61] to obtain the contextualized representations $\mathbf{H}_w = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$ for the text, where each token's representation $\mathbf{h}_j$ is the average of the representations of their

corresponding subtokens. For each token j, we feed its representation $\mathbf{h}_j$ into an event-type classification layer to classify the token into a positive event type or *Other* if it is not an event trigger. Note that all event triggers are single-token.

For argument extraction, we concatenate the contextualized representation of each identified event trigger $\mathbf{h}_{\tau_j}$ with the representation of each argument candidate (i.e., entity mention) $\mathbf{h}_{m_i}$ in $\mathcal{M}$ and feed the concatenated representations into an argument role classification layer to compute the probabilities for argument role types.

Both event detection and argument role classification are optimized with multi-class cross entropy.

JOINT4E-EE. For event extraction, we propose JOINT4E-EE, a dual encoder framework that incorporates the external domain knowledge of the given entities by the base entity linking model such that it can better extract biological processes from unstructured texts. Given the input text $\mathbf{x}^E$ and an entity mention m_i from the set of all entity mentions $\mathcal{M}$, **first** we leverage the search engine of the QuickGO API[1] for GO knowledge base to retrieve a set of candidate biomedical concepts $\mathcal{C}_i$ from the GO knowledge base. We type in the tokens of the entity mention to the search engine and the QuickGO API returns the set of all possible candidates. If there are more than 30 candidates returned, we only take the first 30 candidates returned by the QuickGO API. For the rest of the section, we use the term *retrieve candidate concepts* to refer to the same process mentioned above. Table 3.2 shows the fraction of mentions that can be found with at least one positive candidate. When we take more than 30 candidates, the fraction doesn't increase. **Second**, we apply the base entity linking model to select a biomedical concept from the candidate set of concepts $\mathcal{C}_i$ retrieved from the GO knowledge base. **Third**, we obtain the definition of the corresponding biomedical concept from GO and use it as part of the input for JOINT4E-EE. In particular, we apply an additional

[1] https://www.ebi.ac.uk/QuickGO/

PLM-based knowledge encoder that specifically takes in the selected biomedical definition $\mathbf{d}_i = [d_{i,1}, d_{i,2}, ...d_{i,q}]$ for m_i and encodes it into contextualized representations. We take the contextualized representation of the [CLS] token as the knowledge encoding for m_i, denoted as $\mathbf{v}_i$. We adopt the same process for all the entities in $\mathcal{M}$, which yields a set of knowledge encodings $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p\}$. Meanwhile, similarly to the base event extraction model, we also encode the input text $\mathbf{x}^E$ into contextualized representations $\mathbf{H}_w = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n]$ with a text encoder. **Forth**, we integrate the external knowledge by applying element-wise addition between the contextualized representation of each mention m_i and its corresponding knowledge encoding $\mathbf{v}_i$ such that we obtain a knowledge-enhanced entity representation via $\tilde{\mathbf{h}}_{m_i} = \mathbf{h}_{m_i} + \mathbf{v}_i$. **Finally**, we concatenate the representation of each identified event trigger $\mathbf{h}_{\tau_j}$ with the enhanced entity representation $\tilde{\mathbf{h}}_{m_i}$ and feed it into the classification layer to perform argument extraction.

3.2.2 Event-aware biomedical entity linking

Base Entity Linking Model. The base entity linking model (Base-EL) takes in $\mathbf{x}^L = [x_1^L, x_2^L, ..., x_n^L]$ and the set of spans for all entity mentions $\mathcal{M} = \{m_1, m_2, ..., m_p\}$ in $\mathbf{x}^L$, and maps each entity mention $m_i \in \mathcal{M}$ to a concept in the external knowledge base, i.e., Gene Ontology (GO). We retrieve a set of candidate concepts $\mathcal{C}_i$ from GO for entity mention m_i. For each candidate c_k from the candidate set $\mathcal{C}_i$, we obtain its definition in GO which is also a text sequence, denoted as $\mathbf{d}_k = [d_{k,1}, d_{k,2}, ...d_{k,q}]$. We append the definition $\mathbf{d}_k$ at the end of $\mathbf{x}^L$ separated by a special token [SEP], which yields the whole input sequence for the model:

$$[\text{CLS}][x_1^L, x_2^L, ..., x_n^L][\text{SEP}][d_{k,1}, d_{k,2}, ...d_{k,q}]. \tag{3.1}$$

We encode the entire sequence with a pretrained language model (PLM) encoder [60, 61] and then take the contextualized representation of the [CLS] token output from the encoder to compute the probability $\mathbb{P}(c_k|m_i, \mathbf{x}^L, \mathbf{d}_k; \theta_L)$ with a binary classification layer, where θ_L denotes the parameters of the entity linking model. The model is optimized by the binary cross entropy loss.

JOINT4E-EL We introduce JOINT4E-EL, a framework that utilizes the output of a base event extraction model (see Section 4.1 for details) for biomedical entity linking. JOINT4E-EL consists of a PLM encoder [60, 61] that computes the contextualized representations for the input sequence and a binary classification layer that computes the probability of the mapping between a given entity mention m_i in the input text and a biomedical concept from a candidate set $\mathcal{C}_i$ in the knowledge base. Based on the base entity linking model, we incorporate the event information into the entity linking model to provide more structural context for better entity disambiguation.

Specifically, given the input text $\mathbf{x}^L$, **first**, we apply the base event extraction model to obtain the pseudo trigger and argument role labels. **Second**, we take one entity mention m_i and retrieve candidate concepts $\mathcal{C}_i$ for m_i. **Third**, we inject event information into the input sequence $\mathbf{x}^L$ of the entity linking model. Each entity only participates in a single event τ_i with a unique argument role α_i (if any). We insert the name of the argument role α_i after the tokens of m_i in the original text $\mathbf{x}^L$ and append the name of the event type τ_i at the end of the sentence $\mathbf{x}^L$. Note that we set the name of the argument role as "*Other*" if the entity does not participate in any biological process. Similarly to the base entity linking model, we also append the definition $\mathbf{d}_k$ w.r.t. the candidate concept c_k after the original input. The event-enhanced input sequence for our JOINT4E-EL model is structured as:

$$[\text{CLS}][x_1^L, x_2^L, ..., m_i, \alpha_i, ..., x_n^L, \tau_i][\text{SEP}][d_{k,1}, d_{k,2}, ...d_{k,q}]. \tag{3.2}$$

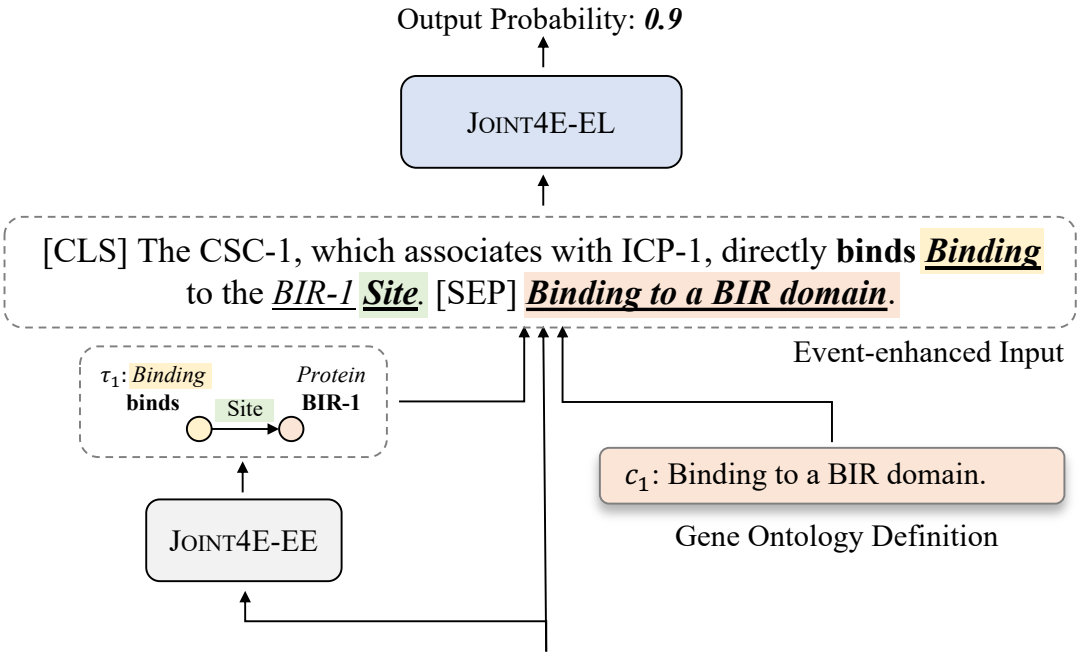

Figure 3.2: Illustration of our Joint4E-EL for biomedical entity linking. Given an entity mention (e.g., *BIR-1*), Joint4E-EL combines the original text, mention definition from Gene Ontology, and the predicted event structure from Joint4E-EE as the event-enhanced input and outputs a probability to indicate its confidence on the candidate concept, e.g., c_1. We select the candidate with the highest probability as the predicted concept.

Forth, We encode the input sequence with the PLM encoder and feed the contextualized representation of the [CLS] token into the binary classification layer.

3.2.3 Iterative collaborative training

In this section, we formulate our iterative collaborative training algorithm which is shown in Algorithm 1. For the event extraction task, we denote a training instance as $(\mathbf{x}_i^E, \mathbf{y}_i^E)$, where $\mathbf{x}_i^E$ is a sentence and $\mathbf{y}_i^E$ is the event annotation on $\mathbf{x}_i^E$, in the event extraction dataset $\mathcal{D}_E$. We denote $\mathcal{Z}_{all}^L$ as the finite set of all possible entity linking labels on the text $\mathbf{x}_i^E$. We further define $\mathcal{Z}^L = \{\mathbf{z}^L \in \mathcal{Z}_{all}^L : f_{\theta_E}(\mathbf{x}_i^E, \mathbf{z}^L) = \mathbf{y}_i^E\}$ as the set of entity linking labels that

leads to the correct event extraction prediction on the sentence $\mathbf{x}_i^E$, where f_{θ_E} is the event extraction model and θ_E denotes its parameters. In our setting, the (pseudo) entity linking labels become discrete latent variables for the event extraction task.

For the entity linking task, given an instance $(\mathbf{x}_i^L, \mathbf{y}_i^L)$ where $\mathbf{x}_i^L$ is a sentence and $\mathbf{y}_i^L$ is the entity linking annotation on $\mathbf{x}_i^L$, in the entity linking dataset $\mathcal{D}_L$, we denote $\mathcal{Z}_{all}^E$ as the finite set of all possible event extraction labels on the text $\mathbf{x}_i^L$. We further define $\mathcal{Z}^E = \{\mathbf{z}^E \in \mathcal{Z}_{all}^E : f_{\theta_L}(\mathbf{x}_i^L, \mathbf{z}^E) = \mathbf{y}_i^L\}$ as the set of event extraction labels that leads to the correct entity linking predictions on the sentence $\mathbf{x}_i^L$, where f_{θ_L} is the entity linking model and θ_L denotes its parameters. In the above setting, the (pseudo) event extraction labels become discrete latent variables for the entity linking task.

Given a dataset $\mathcal{D}_L$ with entity linking annotations and a dataset $\mathcal{D}_E$ with event extraction annotations, we first perform the following prerequisite steps: **First**, We prepare the candidate biomedical concepts for both entity linking dataset $\mathcal{D}_L$ and event extraction dataset $\mathcal{D}_E$. **Second**, we randomly initialize the parameter θ_L for JOINT4E-EE and the parameter θ_E for JOINT4E-EL. To first obtain a well-initialized base model for each task, we individually train JOINT4E-EL on the labeled entity linking dataset $\mathcal{D}_L$ and train JOINT4E-EE on the labeled entity linking dataset $\mathcal{D}_L$ until the model converges on the development sets, respectively. After we obtain a base model individually trained on each task, we start our iterative training process that repeatedly performs the following two collaborative training steps: (1) the Estimation step that aims to estimate the latent variables (i.e., predict pseudo labels) for each partially annotated dataset, and; (2) Updation step where it updates the parameters of each model given the original inputs and the estimated latent variables.

Estimation Step. At the beginning of each round of the iterative training, we first initialize two empty sets $\mathcal{U}_L = \{\}$ and $\mathcal{U}_E = \{\}$ for collecting pseudo labeled instances. We run the entity linking model JOINT4E-EL on the event extraction dataset $\mathcal{D}_E$ to generate pseudo

entity linking annotations. Specifically, for each instance in the event extraction dataset $(\mathbf{x}_i^E, \mathbf{y}_i^E) \in \mathcal{D}_E$, we run the Joint4E-EL model and predict the pseudo entity linking labels $\mathcal{Z}^L$. For the event extraction task, we take the latent variable $\tilde{\mathbf{z}}_i^L \in \mathcal{Z}^L$ that has the highest likelihood, i.e., $\tilde{\mathbf{z}}_i^L = \text{argmax}_{\mathbf{z}_j^L \in \mathcal{Z}^L} \mathbb{P}(z_j^L | x_i^E; \theta_L)$. The estimated latent variable $\tilde{\mathbf{z}}_i^L$ together with $\mathbf{x}_i^E$ and $\mathbf{y}_i^E$ form a new instance and is added into $\mathcal{U}_E$. We also run the event extraction model Joint4E-EE on the entity linking dataset $\mathcal{D}_E$ to generate pseudo event extraction annotations. Specifically, for each instance in the entity linking dataset $(\mathbf{x}_i^L, \mathbf{y}_i^L) \in \mathcal{D}_L$, we run the Joint4E-EE model and predict the pseudo event labels $\mathcal{Z}^E$. For the event extraction task, we take the latent variable $\tilde{\mathbf{z}}_i^E \in \mathcal{Z}^E$ that has the highest likelihood, i.e., $\tilde{\mathbf{z}}_i^E = \text{argmax}_{\mathbf{z}_j^E \in \mathcal{Z}^E} \mathbb{P}(z_j^E | x_i^E; \theta_E)$. The estimated latent variable $\tilde{\mathbf{z}}_i^E$ together with $\mathbf{x}_i^L$ and $\mathbf{y}_i^L$ form a new instance and is added into $\mathcal{U}_L$.

Updation Step. For the event extraction task, we loop through the examples $(\mathbf{x}_i^E, \mathbf{y}_i^E, \tilde{\mathbf{z}}_i^L)$ in the newly collected $\mathcal{U}_E$ event extraction dataset enhanced with pseudo entity linking annotations. The Joint4E-EE model f_θ^E optimizes the log-likelihood of the true event extraction label $\mathbf{y}_i^E$ based on the discrete latent variable $\tilde{\mathbf{z}}_i^L$, i.e., the entity linking pseudo label. The loss is computed as $J_E(\theta_E | \mathbf{x}_i^E, \tilde{\mathbf{z}}_i^L) = -\log \mathbb{P}(\mathbf{y}_i^E | \mathbf{x}_i^E, \tilde{\mathbf{z}}_i^L; \theta_E)$. For the entity linking task, we loop through the examples $(\mathbf{x}_i^L, \mathbf{y}_i^L, \tilde{\mathbf{z}}_i^E)$ in the newly collected $\mathcal{U}_L$ entity linking dataset enhanced with pseudo-event extraction annotations. The Joint4E-EL model f_θ^L optimizes the log-likelihood of the true entity linking label $\mathbf{y}_i^L$ based on the discrete latent variable $\tilde{\mathbf{z}}_i^E$, i.e., the event pseudo label. The loss is computed as $J_L(\theta_L | \mathbf{x}_i^L, \tilde{\mathbf{z}}_i^E) = -\log \mathbb{P}(\mathbf{y}_i^L | \mathbf{x}_i^L, \tilde{\mathbf{z}}_i^E; \theta_L)$.

3.3 Experimental setup

3.3.1 Datasets

Event Extraction

We evaluate the performance of our approach for biomedical event extraction on the Genia 2011 dataset (GE11) [48], which defines 9 event types with 6 argument roles. The text is based on the abstracts and full articles from PubMed about biological processes related to proteins and genes. The detailed statistics of GE11 are summarized in Table 3.1. Following previous studies [43, 55, 56, 57, 62, 63], we evaluate the performance of biomedical event extraction using the precision (P), recall (R), and F1 score (F1). The best model is always validated and selected from the development set and directly evaluated on the test set.

Entity Linking

For the entity linking task, we leverage the BioCreative IV GO (BC4GO) dataset [35] which contains annotations of Gene Ontology entities for all the entity mentions in the dataset, i.e., each entity mention in BC4GO is mapped to a unique biomedical entity in the Gene Ontology knowledge base where each entity is described with GO id, name, and definition. However, the original BC4GO dataset was built in 2013. With the development of Vivo and Vitro in biomedical science in the last decades, new definitions and ontologies of biomedical concepts have been introduced into the Gene Ontology knowledge base, which drastically changes the topology of the knowledge base [64, 65] and makes the mappings between the entity mentions and their concepts in the original BC4GO outdated. In addition, previous studies [59] also suggest that the mappings between entity mentions and entities in the Gene Ontology knowledge base are not surjective, i.e., each entity mention can be mapped into

multiple entities. Thus, we propose to update the mappings between entity mentions in BC4GO and entities in Gene Ontology by leveraging the official API [2] of Gene Ontology.

Specifically,

Gene Ontology consists of three directed acyclic graphs (DAG) and there is no connected component between them. All the nodes in the three DAGs are biomedical concepts and each DAG has a hierarchical structure called an *ancestor chart*. The biomedical concepts at the upper level of an *ancestor chart* have broader meanings compared with the biomedical concepts at the lower level. The three nodes at the top of the three *ancestor chart*s in Gene Ontology are: *Cellular Component*, *Molecular Function*, and *Biological Process*, which are the broadest concepts in biomedical science and any biomedical concept is contained in one of them. We refer to the three topmost nodes in the three *ancestor chart*s as root nodes in the rest of the section. For each node in an *ancestor chart*, the nodes above it are broader concepts while the nodes below it are finer concepts, and there is no edge between the nodes from the same level.

For each entity mention m_i from a sentence $\mathbf{w}$ of the BC4GO dataset, we denote the annotated target entity from Gene Ontology for m_i in the original BC4GO dataset as c_i^{gold}. We also take m_i as a query to the search engine via the API of Gene Ontology and obtain a set of candidate entities $\mathcal{C}_i = \{c_{i1}, c_{i2}, ..., c_{ij}\}$. Then, for each $c_{ij} \in \mathcal{C}_i$, we find a path between c_{ij} and the root node of a particular *ancestor chart*, denoted as $\mathbf{p}_{ij}$. Following a similar process, we also find the path $\mathbf{p}_{gold}$ between the gold target entity c_i^{gold} and a root node. Based on these paths, we design the following four strategies to determine a new set of gold entities from $\mathcal{C}_i$ for each mention m_i as the reference in Gene Ontology:

- If a candidate entity $c_{ij} \in \mathcal{C}_i$ for the mention m_i is the same as the ground-truth

[2] https://www.ebi.ac.uk/QuickGO/

concept c_i^{gold} in the original BC4GO dataset, we add c_{ij} to the gold target entity set.

- If a candidate entity $c_{ij} \in \mathcal{C}_i$ for the mention m_i is on the path $\mathbf{p}_{gold}$, we add c_{ij} to the gold target entity set.
- If the gold target entity c_i^{gold} annotated in the original BC4GO dataset is on the path $\mathbf{p}_{ij}$ for a candidate entity c_{ij}, we add c_{ij} to the gold target entity set.
- If the path $\mathbf{p}_{ij}$ for a particular candidate entity c_{ij} have more than 4 overlapped nodes with the path $\mathbf{p}_{gold}$ for the original gold entity, we add c_{ij} to the gold target entity set.

In order to enhance the robustness of the new mapping method for GO terms, we increase the potential candidate number to let mentions have at least one positive GO term as much as possible (Table 3.2). After setting 30 potential candidates, the expanded BC4GO dataset contains 29,037 mention-candidate pairs in the training set (9,027 positive and 20,010 negative pairs), 7,023 pairs in the dev set (2,352 positive and 4,671 negative pairs), and 5,580 pairs in the test set (1,578 positive and 4,002 negative pairs). During the evaluation, we set each mention with one candidate as one pair and use accuracy to calculate the correct prediction pair number over the total number of mention-candidate pairs.

3.3.2 Baselines

Entity Linking Very few previous studies on biomedical entity linking have focused on the BC4GO dataset while the popularly used entity linking benchmark datasets, such as BC5CDR [66], NCBI [67] and COMETA [68], focus on different entity types, such as *disease*, *chemicals* and *colloquial* terms, which are not related to event extraction. To set up baselines, we adapt several state-of-the-art methods, including LATTE [31], Bootleg [69], Fast Dual Encoder [33] to the BC4GO dataset and compare them with our JOINT4E-EL.

Event Extraction We evaluate the effectiveness of our Joint4E-EE on the Genia 2011 Event Extraction benchmark dataset which is the main task in the BioNLP Shared Task series and defines 9 fine-grained event types, such as *Gene expression, Transcription, Protein catabolism, Phosphorylation, Localization, Binding, Regulation, Positive regulation, and Negative regulation*. We compare Joint4E-EE with several recent state-of-the-art methods on biomedical event extraction, including: TEES [70], EventMine [71], Stacked generalization [72], TEES-CNN [73], KB-driven Tree-LSTM [42], QA with BERT [57], GEANet [43], BEESL [56], DeepEventMine [55], HANN [62], and CPJE [63].

3.3.3 Implementation details

We first train the base event extraction and entity linking models on Genia 2011 and BC4GO datasets, respectively. For the base event extraction model, we use AdamW as the optimizer and train it for 30 epochs with a learning rate 5e-5. The batch size is set to 16. For the base entity linking model, we also use AdamW as the optimizer and train it for 30 epochs. The learning rate is 3e-5 and the batch size is set as 16. We stop the training of these two base models if they do not show a better performance for 5 consecutive epochs. For joint training (both Joint4E-EE and Joint4E-EL), we use learning rates 2e-5, 1e-5, 1e-5, 5e-6, 5e-6, 5e-6 for 6 rounds respectively, and the batch size is set to 16. We stop the training of these two Joint4E-EE and Joint4E-EL models if they do not show a better performance for 5 consecutive epochs.

3.4 Main results

3.4.1 Event extracton

Table 3.3 shows the results of our approach JOINT4E-EE as well as all the strong baselines on Genia 2011 dataset. As we can see, JOINT4E-EE significantly outperforms all the strong baselines on recall and the overall F1 score, demonstrating the effectiveness of our join learning framework. Compared with the baselines such as Tree-LSTM [42] and GEANet [43], which share similar ideas as our approach: both of them disambiguate the biomedical terms in the context by linking them to entities in an external knowledge base and incorporate the entity knowledge into the event extraction model. However, they suffer from the errors that are propagated from the static entity linking step to event extraction while our approach joint improves both JOINT4E-EE and JOINT4E-EL. Several recent studies, such as DeepEventMine [55], HANN [62], and CPJE [63], achieve much higher precision but lower recall compared with JOINT4E-EE.

3.4.2 Entity linking

Table 3.4 shows the performance of various approaches for biomedical entity linking based on the test set of BC4GO. We observe JOINT4E-EL significantly outperforms the three strong baselines. Our base entity linking model shares a similar architecture as [74], however, by incorporating the additional event features from the local context, the accuracy of JOINT4E-EL is improved by a large margin, demonstrating the benefit of event-based features to entity linking.

Algorithm 1: Iterative Collaborative Training for JOINT4E

Input: Entity linking dataset $\mathcal{D}_L$, event extraction dataset $\mathcal{D}_E$, external knowledge base $\mathcal{B}$, learning rates η_L and η_E.

for *each entity set* $\mathcal{M}$ *in* $\mathcal{D}_L$ **do**
  for $m_i \in \mathcal{M}$ **do**
    Retrieve the candidate set $\mathcal{C}_i$ for m_i from $\mathcal{B}$;
  end
end
Initialize the entity linking model's parameters θ_L and the event extraction model's parameters θ_E;
Train θ_L on $\mathcal{D}_L$ and θ_E on $\mathcal{D}_E$;
`// Collaborative learning`
while *not converged* **do**
  `// Initialize augmented datasets`
  $\mathcal{U}_L = \{\}, \mathcal{U}_E = \{\}$;
  `// Estimation step`
  for *each* $(\mathbf{x}_i^L, \mathbf{y}_i^L) \in \mathcal{D}_L$ **do**
    $\tilde{\mathbf{z}}_i^E = \text{argmax}_{\mathbf{z}_j^E \in \mathcal{Z}^E} \mathbb{P}(\mathbf{z}_j^E | \mathbf{x}_i^L; \theta_E)$;
    $\mathcal{U}_L \leftarrow \mathcal{U}_L \bigcup \{(\mathbf{x}_i^L, \mathbf{y}_i^L, \tilde{\mathbf{z}}_i^E)\}$;
  end
  for *each* $(\mathbf{x}_i^E, \mathbf{y}_i^E) \in \mathcal{D}_E$ **do**
    $\tilde{\mathbf{z}}_i^L = \text{argmax}_{\mathbf{z}_j^L \in \mathcal{Z}^L} \mathbb{P}(z_j^L | x_i^E; \theta_L)$;
    $\mathcal{U}_E \leftarrow \mathcal{U}_E \bigcup \{(\mathbf{x}_i^E, \mathbf{y}_i^E, \tilde{\mathbf{z}}_i^L)\}$;
  end
  `// Updation step`
  for *each epoch* **do**
    Sample $(\mathbf{x}_i^L, \mathbf{y}_i^L, \tilde{\mathbf{z}}_i^E) \sim \mathcal{U}_L$;
    $\theta_L \leftarrow \theta_L - \eta_L \nabla_{\theta_L} J_L(\theta_L | \mathbf{x}_i^L, \tilde{\mathbf{z}}_i^E)$;
  end
  for *each epoch* **do**
    Sample $(\mathbf{x}_i^E, \mathbf{y}_i^E, \tilde{\mathbf{z}}_i^L) \sim \mathcal{U}_E$;
    $\theta_E \leftarrow \theta_E - \eta_E \nabla_{\theta_E} J_E(\theta_E | \mathbf{x}_i^E, \tilde{\mathbf{z}}_i^L)$;
  end
end

GE11	Training	Development	Test
# Documents	908	259	347
# Sentences	8,664	2,888	3,363
# Entities	11,625	4,690	5,301
# Events	10,310	3,250	4,487

Table 3.1: Statistics of the Genia 2011 dataset for biomedical event extraction.

GO number	10	15	20	25	**30**	35
Fraction	0.56	0.68	0.79	0.82	0.83	0.83

Table 3.2: The fractions of mentions that can be found with at least one positive candidate.

Method	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
TEES [70]	57.65	49.56	53.30
EventMine [71]	63.48	53.35	57.98
Stacked generalization [72]	66.46	48.96	56.38
TEES-CNN [73]	69.45	49.94	58.10
KB-driven Tree-LSTM [42]	67.01	52.14	58.65
QA with BERT [57]	59.33	57.37	58.33
GEANet [43]	64.61	56.11	60.06
BEESL [56]	69.72	53.00	60.22
DeepEventMine [55]	70.52	56.52	62.75
HANN [62]	71.73	53.21	61.10
CPJE [63]	**72.62**	53.33	61.50
Base-EE (Ours)	68.20	55.73	61.23
JOINT4E-EE (Ours)	69.66	**59.75**	**64.35**

Table 3.3: Performance comparison of various event extraction approaches on the *development* set of BioNLP Genia 2011. (%). Bold highlights the highest performance among all the approaches.

Method	**Accuracy (%)**
LATTE [31]	82.71
Bootleg [69]	78.51
Fast Dual Encoder [33]	82.03
Base-EL (Ours)	81.35
Joint4E-EL (Ours)	**85.08**

Table 3.4: Performance comparison of various entity linking approaches on the *test* set of BC4GO. Bold highlights the highest performance among all the approaches.

Chapter 4

Discussion and Conclusion

4.1 Impact of the number of training rounds

Tables 4.1 show the performance of both event extraction and entity linking at each round of joint training based on the iterative collaborative algorithm. We observe that the performance of both models gradually increases with more rounds of joint training and both models achieve the highest performance after 3 rounds. Compared with the base models, both JOINT4E-EE and JOINT4E-EL achieve significant improvements with a large marge: 3.12% absolute F1 score gain for event extraction and 3.73% absolute accuracy gain for entity linking, demonstrating the effectiveness of our joint learning framework. Table 4.2 shows the event extraction performance (i.e., F1 Score) on each fine-grained event type and three event type categories (including simple events, binding events and complex events) at each round during joint training. As we can see, with 3-4 rounds of joint training, JOINT4E-EE achieves up to 2.26%, 6.17%, and 2.99% absolute F1 score gain on the simple, binding and complex events, indicating that binding events benefit the most from the entity knowledge from external knowledge bases. This is consistent with our observation as many entity descriptions in the knowledge base indicate the binding functions of the entities. We also observe that, with more rounds of joint training, the performance of JOINT4E-EE decreases more on complex events which contain multiple arguments and nested events, such as *regulation*, *positive regulation*, and *negative regulation*.

	Event extraction	Entity linking
Rounds	F1 Score(%)	Accuracy(%)
Base	61.23	81.35
1st	62.48	83.38
2nd	63.85	84.06
3rd	**64.35**	**85.08**
4th	64.28	**85.08**
5th	64.15	**85.08**
6th	64.15	**85.08**

Table 4.1: Results of event extraction on the Genia 2011 development set and entity linking on the test set of BC4GO at each round of joint training. Bold highlights the highest performance among all the approaches.

4.2 Qualitative analysis

Table 4.3 shows three examples for which the event predictions are improved and corrected within the first 3 rounds of joint training. Taking the first sentence as an example, before the first round of joint training, JOINT4E-EE mistakenly predicts a *Phosphorylation* event triggered by "phospho" with "STAT3" as the *Theme* argument due to the misinterpretation of the sentence. However, the entity knowledge retrieved from the Gene Ontology (GO) by JOINT4E-EL indicates that "STAT3 is a regulation of tyrosine STAT protein and BMP-6 is a regulation of BMP signaling pathway", while the word "regulation" from both GO definitions helps better disambiguate the context during the 1st round of joint training and finally calibrates the previous wrong event predictions to the *Regulation* event triggered by "changes" with two arguments: "STAT3" as the *Theme* and "BMP-6 as the *Cause* argument.

Similarly, Table 4.4 also shows three examples for which the entity linking results are improved and corrected within the first 3 rounds of joint training. Taking the first sentence as an example, before the 1st round of joint training, JOINT4E-EL mistakenly links the entity mention "*UNC-75*" to the entity defined by "*Positive regulation of synaptic transmission*".

Event type	Base F1 Score	Round 1 F1 Score	Round 2 F1 Score	Round 3 F1 Score	Round 4 F1 Score	Round 5 F1 Score	Round 6 F1 Score
Gene expression	78.15	79.08	80.23	81.03	81.12	81.00	81.00
Transcription	69.46	70.60	71.14	72.08	72.08	72.08	72.08
Protein catabolism	74.57	75.06	75.88	76.38	76.38	76.38	76.38
Phosphorylation	83.67	84.12	84.95	85.67	85.67	85.67	85.67
Localization	80.30	80.75	81.07	81.30	81.30	81.30	81.30
Simple events	76.73	77.92	78.65	78.79	**78.99**	78.77	78.77
Binding	52.19	55.26	56.73	**58.36**	**58.36**	**58.36**	**58.36**
Regulation	45.52	46.03	47.94	49.03	48.53	48.53	48.53
Positive regulation	49.52	50.82	52.02	52.95	52.52	52.41	52.41
Negative regulation	57.48	57.78	58.09	58.52	58.33	58.33	58.33
Complex events	50.84	51.54	52.68	**53.83**	53.42	53.31	53.31

Table 4.2: F1 scores (%) of event extraction on the Genia 2011 develop set for each fine-grained event type and three categories (simple events, binding events, and complex events) at each round. Bold highlights the highest performance among all the approaches.

However, by incorporating the event knowledge with joint training, especially knowing that "*UNC-75*" is the *Theme* of a *Binding* event, JOINT4E-EL correctly links "*UNC-75*" to the target entity defined by "*single-stranded RNA binding*" in the Gene Ontology.

4.3 Error analysis

We further sample 50 prediction errors for both entity linking and event extraction based on their results on the development set of each dataset, respectively. We summarize the main error categories for each task as follows:

Entity Linking: 76% (38/50) of the remaining error for entity linking lies in the candidate retrieval where the candidate sets retrieved based on the Gene Ontology (GO) API for some entity mentions do not contain their true target entities. For example, for the entity mention "*TAT-DeltaDBD-GATA3*", the candidate set returned by GO API does not include the true target entity GO:0019799 with name *acetyl-CoA:alpha-tubulin-L-lysine 6-N-acetyltransferase*

activity.

Event Extraction: The main error (33/50) for event extraction lies in the missing or spurious argument predictions. Most event types such as simple events (including *gene expression*, *transcription*, *localization*, *phosphorylation*, and *protein catabolism*) are defined with a fixed number of arguments while the complex events and binding events are usually associated with up to four possible arguments, thus the model tends to miss some arguments or predict spurious arguments. Taking the following two sentences as examples:

- **S1:** The **FOXP3** (***arg: Theme***) **inhibition** (***trigger: Negative regulation***) by GATA element in the FOXP3 **promoter** **(redundant arg: Site)**.

- **S2:** Disruption of the **Jak1** (***arg: Theme***) **binding** (***trigger: Binding***), proline-rich **Box1** (***arg: Site***) region of **IL-4R** **(missing arg: Theme)** abolished signaling by this chimeric receptor.

For S1, our model successfully predicts *inhibition* as a Negative regulation event and *FOXP3* as its *Theme* argument. However, it also mistakenly predicts *promoter* as a *Site* argument, due to two possible reasons: (1) the entity *promoter* is frequently labeled as a *Site* argument in the training set; and (2) the protein *FOXP3* is defined as "*regulation of DNA-templated transcription*" in the Gene Ontology, which also tends to imply *promoter* as a *Site* argument. In S2, our Joint4E-EE correctly predicts *binding* as a *Binding* event with two arguments: *Jak1* and *Box1*. However, it mistakenly misses another *Theme* argument which is likely because the model treats *IL-4R* as *Box1* which is already labeled as a *Site* argument.

Rounds	
1st	**Text:** We did not observe any significant **changes** in the level of **phospho**-*STAT3* or phospho-p38 upon *BMP-6* treatment of B cells.
	Previous Results: **Event type**: Phosphorylation; **Trigger**: phospho; **Theme**: STAT3.
	Entity Knowledge from GO: **STAT3**: regulation of tyrosine STAT protein; **BMP-6**:regulation of BMP signaling pathway.
	New Results: **Event type**: Regulation; **Trigger**: changes; **Theme**: STAT3, **Cause**: BMP-6.
2nd	**Text:** Costimulation through *CD28* and/or CD2 did not **modulate**, the CD3-dependent **phosphorylation** of HS1.
	Previous Results: **Event type**: Phosphorylation; **Trigger**: phosphorylation; **Theme**: CD28.
	Entity Knowledge from GO: **CD28**: immune response; **CD2**: regulation of CD4, CD25 regulatory T cell differentiation.
	New Results: **Event type**: Regulation; **Trigger**: modulate; **Theme**: phosphorylation (event), **Cause**: CD28.
3rd	**Text:** When tested its ability to block calcineurin-dependent signaling in cells, the pivotal promoter element for *interleukin-2* gene **induction**.
	Previous Results: **Event type**: Regulation; **Trigger**: induction; **Theme**: interleukin-2.
	Entity Knowledge from JOINT4E-EL: **interleukin-2**: plastid gene expression.
	New Results: **Event type**: Gene expression; **Trigger**: induction; **Theme**: interleukin-2.

Table 4.3: Example sentences and results for event extraction at each round sampled from the Genia 2011 development set. For each sentence, before each round of joint training, the event prediction is not correct while after incorporating the entity knowledge from JOINT4E-EL, the errors are corrected with joint training. The bold words in each text highlight the candidate event triggers while the italic words highlight the candidate arguments predicted by JOINT4E-EE.

4.4 Conclusion and Future work

In this work, we propose a joint biomedical entity linking and event extraction framework, i.e., JOINT4E-ELand JOINT4E-EE, to leverage the benefit of one task to the other. Our JOINT4E-EEcan incorporate the domain knowledge obtained by JOINT4E-EL, while JOINT4E-EL can be improved by the event structural context provided by JOINT4E-EE. To iteratively improve the two tasks together, we propose a novel iterative collaborative training strategy where we first estimate missing variables for both two incomplete datasets based on the current task-specific models, and then update the parameters of both models on the datasets that are augmented by pseudo labels. We conduct extensive experiments on the biomedical entity linking dataset, i.e., BC4GO, and biomedical event extraction, i.e., Genia 11. We also provide several valuable discussions such as error analysis that reveals the remaining challenges of both two tasks. We hope this work can shed light on the following research on biomedical information extraction and broader communities. In the future, we plan to expand our framework to other datasets across other biomedical domains such as diseases,

Rounds	
1st	**Text:** To determine the elements in the exon 7 region that ***UNC-75*** directly and specifically recognizes in vitro. **Previous Results:** **Entity:** Positive regulation of synaptic transmission. **Event Knowledge from Joint4E-EE: Event type**: Binding; **Trigger**: recognize; **Theme**: UNC-75. **New Results:** **Entity:** single-stranded RNA binding.
2nd	**Text:** These data suggest that ectopically expressed ***BP*** downregulates FIL activity genes in ovaries. **Previous Results:** **Entity:** DNA-binding transcription factor activity, RNA polymerase II-specific. **Event Knowledge from Joint4E-EE: Event type**: negative regulation; **Trigger**: downregulates; **Theme**: FIL. **New Results:** **Entity:** negative regulation of gene expression.
3rd	**Text:** Effects of the ***gei-8*** mutation on gene expression were studied with whole genome microarrays. **Previous Results:** **Entity:** reciprocal meiotic recombination. **Event Knowledge from Joint4E-EE: Event type**: gene expression; **trigger**: expression; **Theme**: gei-8. **New Results:** **Entity:** transcription by RNA polymerase II.

Table 4.4: Example sentences and results for entity linking at each round sampled from the development set of BC4GO dataset. For each sentence, before each round of joint training, the entity linking result is not correct while after incorporating the event knowledge from Joint4E-EE, the errors are corrected with joint training. The bold words in each text highlight the candidate entity mention for entity linking.

drugs, and chemicals, in order to test the generalizability and adaptation of our framework. Recently, contrastive learning has shown promising results in various domains and we will try contrastive learning to see whether the performance of both tasks can be enhanced as well.

Bibliography

[1] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, M. Zottoli, M. Nazarzadeh, *et al.*, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *Journal of biomedical informatics*, vol. 101, p. 103337, 2020.

[2] M. AlShuweihi, S. A. Salloum, and K. Shaalan, "Biomedical corpora and natural language processing on clinical text in languages other than english: a systematic review," *Recent Advances in Intelligent Systems and Smart Applications*, pp. 491–509, 2021.

[3] D. S. Bitterman, T. A. Miller, R. H. Mak, and G. K. Savova, "Clinical natural language processing for radiation oncology: a review and practical primer," *International Journal of Radiation Oncology* Biology* Physics*, vol. 110, no. 3, pp. 641–655, 2021.

[4] Z. Liu, R. A. Roberts, M. Lal-Nag, X. Chen, R. Huang, and W. Tong, "Ai-based language models powering drug discovery and development," *Drug Discovery Today*, vol. 26, no. 11, pp. 2593–2607, 2021.

[5] J. L. Snoep, "The silicon cell initiative: working towards a detailed kinetic description at the cellular level," *Current opinion in biotechnology*, vol. 16, no. 3, pp. 336–343, 2005.

[6] R. Randhawa, C. Shaffer, and J. Tyson, "Model composition for macromolecular regulatory networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 278–287, 2008.

[7] K. D. Brumfield, P. Cox, J. Geyer, and J. Goepp, "A taxonomy-agnostic approach to

targeted microbiome therapeutics—leveraging principles of systems biology," *Pathogens*, vol. 12, no. 2, p. 238, 2023.

[8] J. Sung, S. Kim, J. J. T. Cabatbat, S. Jang, Y.-S. Jin, G. Y. Jung, N. Chia, and P.-J. Kim, "Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis," *Nature communications*, vol. 8, no. 1, p. 15393, 2017.

[9] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models—a review," *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[10] J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks," *PLoS computational biology*, vol. 6, no. 6, p. e1000807, 2010.

[11] M. AY, K.-I. Goh, M. E. Cusick, A.-L. Barabasi, M. Vidal, *et al.*, "Drug–target network," *Nature biotechnology*, vol. 25, no. 10, pp. 1119–1127, 2007.

[12] Y. Kim, Y.-S. Jung, J.-H. Park, S.-J. Kim, and Y.-R. Cho, "Drug-disease association prediction using heterogeneous networks for computational drug repositioning," *Biomolecules*, vol. 12, no. 10, p. 1497, 2022.

[13] C. W. Lee, S. M. Kim, S. Sa, M. Hong, S.-M. Nam, and H. W. Han, "Relationship between drug targets and drug-signature networks: a network-based genome-wide landscape," *BMC Medical Genomics*, vol. 16, no. 1, p. 17, 2023.

[14] P. O. Williamson and C. I. Minter, "Exploring pubmed as a reliable resource for scholarly communications services," *Journal of the Medical Library Association: JMLA*, vol. 107, no. 1, p. 16, 2019.

[15] I. Shakeel, N. Basheer, G. M. Hasan, M. Afzal, and M. I. Hassan, "Polo-like kinase 1 as an emerging drug target: structure, function and therapeutic implications," *Journal of Drug Targeting*, vol. 29, no. 2, pp. 168–184, 2021.

[16] I. H. Ibrahim, A. Balah, A. G. A. E. Hassan, and H. G. Abd El-Aziz, "Role of motor proteins in human cancers," *Saudi Journal of Biological Sciences*, p. 103436, 2022.

[17] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of biocreative: critical assessment of information extraction for biology," 2005.

[18] S. Zhao, C. Su, Z. Lu, and F. Wang, "Recent advances in biomedical literature mining," *Briefings in Bioinformatics*, vol. 22, no. 3, p. bbaa057, 2021.

[19] J. S. Amberger and A. Hamosh, "Searching online mendelian inheritance in man (omim): a knowledgebase of human genes and genetic phenotypes," *Current protocols in bioinformatics*, vol. 58, no. 1, pp. 1–2, 2017.

[20] O. Al-Harazi, A. El Allali, and D. Colak, "Biomolecular databases and subnetwork identification approaches of interest to big data community: an expert review," *Omics: a journal of integrative biology*, vol. 23, no. 3, pp. 138–151, 2019.

[21] L. Amos, D. Anderson, S. Brody, A. Ripple, and B. L. Humphreys, "Umls users and uses: a current overview," *Journal of the American Medical Informatics Association*, vol. 27, no. 10, pp. 1606–1611, 2020.

[22] R. Harpaz, A. Callahan, S. Tamang, Y. Low, D. Odgers, S. Finlayson, K. Jung, P. LePendu, and N. H. Shah, "Text mining for adverse drug events: the promise, challenges, and state of the art," *Drug safety*, vol. 37, pp. 777–790, 2014.

[23] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: man-

agement, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 1, pp. 1–25, 2019.

[24] R. Wojciechowski, "Nature and nurture: the complex genetics of myopia and refractive error," *Clinical genetics*, vol. 79, no. 4, pp. 301–320, 2011.

[25] A. T. Young, D. Amara, A. Bhattacharya, and M. L. Wei, "Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review," *The Lancet Digital Health*, vol. 3, no. 9, pp. e599–e611, 2021.

[26] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of bionlp'09 shared task on event extraction," in *Proceedings of the BioNLP 2009 workshop companion volume for shared task*, pp. 1–9, 2009.

[27] F. Leitner, S. A. Mardis, M. Krallinger, G. Cesareni, L. A. Hirschman, and A. Valencia, "An overview of biocreative ii. 5," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 385–399, 2010.

[28] I. Segura-Bedmar, P. Martínez Fernández, and M. Herrero Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," Association for Computational Linguistics, 2013.

[29] C. N. Arighi, P. M. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-Aryamontri, S. Clematide, P. Gaudet, M. G. Giglio, I. Harrow, *et al.*, "Biocreative iii interactive task: an overview," *BMC bioinformatics*, vol. 12, no. 8, pp. 1–21, 2011.

[30] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 183–191, 2011.

[31] M. Zhu, B. Celikkaya, P. Bhatia, and C. K. Reddy, "Latte: Latent type modeling

for biomedical entity linking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9757–9764, 2020.

[32] R. Angell, N. Monath, S. Mohan, N. Yadav, and A. McCallum, "Clustering-based inference for biomedical entity linking," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Online), pp. 2598–2608, Association for Computational Linguistics, June 2021.

[33] R. Bhowmik, K. Stratos, and G. de Melo, "Fast and effective biomedical entity linking using a dual encoder," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, (online), pp. 28–37, Association for Computational Linguistics, Apr. 2021.

[34] G. O. Consortium, "Gene ontology annotations and resources," *Nucleic acids research*, vol. 41, no. D1, pp. D530–D535, 2012.

[35] K. Van Auken, M. L. Schaeffer, P. McQuilton, S. J. Laulederkind, D. Li, S.-J. Wang, G. T. Hayman, S. Tweedie, C. N. Arighi, J. Done, *et al.*, "Bc4go: a full-text corpus for the biocreative iv go task," *Database*, vol. 2014, 2014.

[36] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[37] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D154–D159, 2005.

[38] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne,

A. van den Broek, M. Castro, G. Cochrane, *et al.*, "The embl nucleotide sequence database," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D29–D33, 2005.

[39] M. Čuljak, A. Spitz, R. West, and A. Arora, "Strong heuristics for named entity linking," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, (Hybrid: Seattle, Washington + Online), pp. 235–246, Association for Computational Linguistics, July 2022.

[40] L. Chen, G. Varoquaux, and F. M. Suchanek, "A lightweight neural model for biomedical entity linking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 12657–12665, 2021.

[41] G. Vader, R. H. Medema, and S. M. Lens, "The chromosomal passenger complex: guiding aurora-b through mitosis," *The Journal of cell biology*, vol. 173, no. 6, pp. 833–837, 2006.

[42] D. Li, L. Huang, H. Ji, and J. Han, "Biomedical event extraction based on knowledge-driven tree-lstm," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1421–1430, 2019.

[43] K.-H. Huang, M. Yang, and N. Peng, "Biomedical event extraction with hierarchical knowledge graphs," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 1277–1285, Association for Computational Linguistics, Nov. 2020.

[44] H. Fei, Y. Ren, Y. Zhang, D. Ji, and X. Liang, "Enriching contextualized language model from knowledge graph for biomedical information extraction," *Briefings in bioinformatics*, vol. 22, no. 3, p. bbaa110, 2021.

[45] Z. Zhang, N. N. Parulian, H. Ji, A. S. Elsayed, S. Myers, and M. Palmer, "Fine-grained information extraction from biomedical literature based on knowledge-enriched abstract meaning representation," in *Proc. The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

[46] T. Lai, H. Ji, C. Zhai, and Q. H. Tran, "Joint biomedical entity and relation extraction with knowledge-enhanced collective inference," *arXiv preprint arXiv:2105.13456*, 2021.

[47] M. Van Nguyen, B. Min, F. Dernoncourt, and T. Nguyen, "Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4363–4374, 2022.

[48] J.-D. Kim, Y. Wang, T. Takagi, and A. Yonezawa, "Overview of Genia event task in BioNLP shared task 2011," in *Proceedings of BioNLP Shared Task 2011 Workshop*, (Portland, Oregon, USA), pp. 7–15, Association for Computational Linguistics, June 2011.

[49] L. Wu, F. Petroni, M. Josifoski, S. Riedel, and L. Zettlemoyer, "Scalable zero-shot entity linking with dense entity retrieval," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), pp. 6397–6407, Association for Computational Linguistics, 2020.

[50] M. Varma, L. Orr, S. Wu, M. Leszczynski, X. Ling, and C. Ré, "Cross-domain data integration for named entity disambiguation in biomedical text," *arXiv preprint arXiv:2110.08228*, 2021.

[51] D. Xu, Z. Zhang, and S. Bethard, "A generate-and-rank framework with semantic type regularization for biomedical concept normalization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 8452–8464, Association for Computational Linguistics, July 2020.

[52] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee, "Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition," *arXiv preprint arXiv:2010.03746*, 2020.

[53] R. Han, Q. Ning, and N. Peng, "Joint event and temporal relation extraction with shared representations and structured prediction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Hong Kong, China), pp. 434–444, Association for Computational Linguistics, Nov. 2019.

[54] K. J. Espinosa, M. Miwa, and S. Ananiadou, "A search-based neural model for biomedical nested and overlapping event detection," pp. 3677–3684, 2019.

[55] H.-L. Trieu, T. T. Tran, K. N. 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, 2020.

[56] A. Ramponi, R. van der Goot, R. Lombardo, and B. Plank, "Biomedical event extraction as sequence labeling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5357–5367, 2020.

[57] X. D. Wang, L. Weber, and U. Leser, "Biomedical event extraction as multi-turn question answering," in *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pp. 88–96, 2020.

[58] G. O. Consortium, "Gene ontology annotations and resources," *Nucleic acids research*, vol. 41, no. D1, pp. D530–D535, 2012.

[59] R. Balakrishnan, M. A. Harris, R. Huntley, K. Van Auken, and J. M. Cherry, "A guide to best practices for gene ontology (go) manual annotation," *Database*, vol. 2013, 2013.

[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[61] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[62] 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," *Information Sciences*, vol. 550, pp. 27–40, 2021.

[63] Y. Wang, J. Wang, H. Lu, B. Xu, Y. Zhang, S. K. Banbhrani, H. Lin, *et al.*, "Conditional probability joint extraction of nested biomedical events: Design of a unified extraction framework based on neural networks," *JMIR Medical Informatics*, vol. 10, no. 6, p. e37804, 2022.

[64] Y. R. Park, J. Kim, H. W. Lee, Y. J. Yoon, and J. H. Kim, "Gochase-ii: correcting semantic inconsistencies from gene ontology-based annotations for gene products," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–7, 2011.

[65] S. Yon Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nature Reviews Genetics*, vol. 9, no. 7, pp. 509–515, 2008.

[66] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J.

Mattingly, T. C. Wiegers, and Z. Lu, "Biocreative v cdr task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[67] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: a resource for disease name recognition and concept normalization," *Journal of biomedical informatics*, vol. 47, pp. 1–10, 2014.

[68] M. Basaldella, F. Liu, E. Shareghi, and N. Collier, "Cometa: A corpus for medical entity linking in the social media," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3122–3137, 2020.

[69] L. Orr, M. Leszczynski, S. Arora, S. Wu, N. Guha, X. Ling, and C. Re, "Bootleg: Chasing the tail with self-supervised named entity disambiguation," *arXiv preprint arXiv:2010.10363*, 2020.

[70] J. Björne and T. Salakoski, "Generalizing biomedical event extraction," in *Proceedings of BioNLP Shared Task 2011 Workshop*, (Portland, Oregon, USA), pp. 183–191, Association for Computational Linguistics, June 2011.

[71] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. Tsujii, and S. Ananiadou, "Event extraction across multiple levels of biological organization," *Bioinformatics (Oxford, England)*, vol. 28, p. i575—i581, September 2012.

[72] A. Majumder, A. Ekbal, and S. K. Naskar, "Biomolecular event extraction using a stacked generalization based classifier," in *Proceedings of the 13th International Conference on Natural Language Processing*, (Varanasi, India), pp. 55–64, NLP Association of India, Dec. 2016.

[73] J. Björne and T. Salakoski, "Biomedical event extraction using convolutional neural net-

works and dependency parsing," in *Proceedings of the BioNLP 2018 workshop*, pp. 98–108, 2018.

[74] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using bert in the machine reading comprehension framework," *Journal of Biomedical Informatics*, vol. 118, p. 103799, 2021.