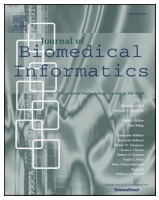
Journal of Biomedical Informatics 139 (2023) 104318

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Joint learning-based causal relation extraction from biomedical literature

Dongling Li [a], Pengchao Wu [a], Yuehu Dong [a], Jinghang Gu [b], Longhua Qian [a,*], Guodong Zhou [a]

[a] School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu Province 215006, China
[b] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong 999077, China

ARTICLE INFO

Keywords:
Joint Learning
BEL Statement
Relation Extraction
Function Detection

ABSTRACT

Causal relation extraction of biomedical entities is one of the most complex tasks in biomedical text mining, which involves two kinds of information: entity relations and entity functions. One feasible approach is to take relation extraction and function detection as two independent sub-tasks. However, this separate learning method ignores the intrinsic correlation between them and leads to unsatisfactory performance. In this paper, we propose a joint learning model, which combines entity relation extraction and entity function detection to exploit their commonality and capture their inter-relationship, so as to improve the performance of biomedical causal relation extraction. Experimental results on the BioCreative-V Track 4 corpus show that our joint learning model outperforms the separate models in BEL statement extraction, achieving the F1 scores of 57.0% and 37.3% on the test set in Stage 2 and Stage 1 evaluations, respectively. This demonstrates that our joint learning system reaches the state-of-the-art performance in Stage 2 compared with other systems.

1. Introduction

With the rapid development of biomedical research, the volume of biomedical literature is growing consistently. When dealing with such large-scale textual data, it is extremely difficult to manually mine useful information from them because of intensive labor and prohibitive cost. Therefore, researchers pay more and more attention to how to automatically obtain relevant knowledge from biomedical literature efficiently and effectively, which leads to the emergence of biomedical text mining research. Biomedical text mining aims to extract useful information from biomedical literature and transform it into structured knowledge to enrich the content of the domain-specific knowledge base. Concurrently, the structured representation of biomedical knowledge is also one of the hot spots pursued by field experts due to its application in knowledge graph construction, knowledge reasoning, and completion. Some well-known attempts for biomedical entity networks are the system biology markup language (SBML) [1], biological pathway exchange language (BioPAX) [2], and biological expression language (BEL) [3].

BEL is a formal language with a specific form used in the biomedical field to represent scientific findings in the life sciences, and it is not only suitable for machine processing but also easily readable for humans. BEL can express complex causal relations between entities and those between entity functions. The BioCreative-V community [4] organized the shared task 4, which aims to extract causal relations between biomedical entities from literature and express them in the BEL form. The extraction of BEL statements poses new research challenges to the biomedical text mining community due to their complex structures. Previous studies either mine biomedical events and transform them into BEL statements [5–9] or extract entity functions/relations and combine them into BEL statements [10–13].

BEL statement extraction involves two sub-tasks: entity relation extraction and entity function detection, it thus is essentially taken as a multi-task learning (MTL) problem. For deep learning models, MTL generally improves learning performance by hard or soft parameter sharing in neural networks. The hard parameter sharing means sharing some hidden layers of all tasks in order to mitigate the risk of overfitting during training (Baxter et al., 1997) [14]. They either merely share the word embeddings (Collobert et al., 2008) [15], or share the whole sentence encoder (Liu et al., 2019 [16]), or generate labels of different tasks at different levels of the neural networks (Søgaard et al., 2016 [17], Sanh et al., 2019 [18]). Soft parameter sharing means that each sub-task has its own model with respective parameters, and the similarity of some parameters is guaranteed by regularizing the distance metric between model parameters. The regularization method can be the L2 norm (Duong et al., 2015 [19]) or the trace norm (Yang et al., 2017 [20]). Multi-task learning in information extraction involving entities,

* Corresponding author.
E-mail addresses: 20204227045@stu.suda.edu.cn (D. Li), 20204227037@stu.suda.edu.cn (P. Wu), 20215227045@stu.suda.edu.cn (Y. Dong), gujinghangnlp@gmail.com (J. Gu), qianlonghua@suda.edu.cn (L. Qian), gdzhou@suda.edu.cn (G. Zhou).

https://doi.org/10.1016/j.jbi.2023.104318
Received 30 June 2022; Received in revised form 3 February 2023; Accepted 8 February 2023
Available online 11 February 2023

Table 1
The overview of the selected BEL functions.

Function	Abbr.	Definition
activities	*act*	The frequency of events resulting from the molecular activity of the proteins, such as the kinase activity of HGNC:AKT1.
complex	*complex*	The abundance of a molecular complex of terms, a list of terms supplied as arguments, such as the protein complex of HGNC:FOS and HGNC:JUN.
protein modification	*pmod*	Covalent modification of the specified protein, such as the phosphorylation of HGNC:AKT1.
degradation	*deg*	The abundance of events in which a term is degraded in some way such that it is no longer the term, such as the proteasome-mediated proteolysis of HGNC:MYC.
cell secretion	*sec*	The abundance of events in which a term moves from cells to regions outside of the cells, such as the secretion of the MGI:IL6 protein.
translocation	*tloc*	The abundance of events in which a term moves from the one location to the other location, such as the endocytosis of the HGNC:EGFR protein.

relations, and events usually adopts hard parameter sharing of token or span representations. Prior works use a sequence labeling model for entity recognition and a tree-based model (Miwa and Bansal, 2016[21]) or a multi-label head selection strategy (Bekoulis et al., 2018[22]) for relation extraction. Recent systems such as DYGIE (Luan et al., 2019 [23]) and DYGIE++ (Wadden et al., 2019[24]) propose to learn and update shared span representations among sub-tasks of entity recognition, relation extraction and coreference resolution via dynamic graph propagation. ONEIE (Lin et al., 2020[25]) further extends DYGIE++ by incorporating global features to capture interactions across sub-tasks (i.e., entity recognition, relation extraction and event extraction) and across instances.

Motivated by the success of multi-task learning in NLP and SBEL-based causality extraction [13], we propose a joint learning model of relation extraction and function detection based on hard parameter sharing [16] for BEL statement extraction. These two sub-tasks share the same BERT encoder but with different fully connected layers and output layers. The difference between our joint model and the aforementioned joint models in information extraction is that they usually deal with entity recognition and relation extraction, while ours focuses on the relation between two given entities and their respective functions (they can also be regarded as unary relations or properties). Our joint model is trained with a novel joint loss function. Specifically, different function types in the loss function are assigned with different weights, which improves the function detection precision by penalizing the negative instances and further promotes the overall extraction performance for BEL statements.

2. Related work

In terms of the key element to be focused on for extraction, the methods used to extract BEL statements can be roughly divided into two groups: event-centered and entity-centered.

Thanks to many publicly available biomedical event corpora, the event-centered approach aims to map biomedical events extracted from literature to entity functions and relations and then formulate them into BEL statements. Ravikumar et al., 2017 [5] propose to extract entity and event information through a rule-based semantic analyzer [6] and then combine them into a complete BEL statement. The disadvantages of rule-based methods are low coverage and poor generalization ability to other domains. Choi et al., 2016 [7] use the event extraction system of machine learning-based TEES [8] to extract biomedical events and then convert them into entity relations and functions in BEL statements. They also discuss the impact of entity co-reference resolution on the extraction performance of BEL statements. Although there is much similarity between BEL statements and biomedical events [9], task difference between them is non-trivial, leading to unsatisfactory performance for BEL statements.

The entity-centered approach takes entities as the core component, and directly determines entity functions and causal relations between entities, and then formulates them into BEL statements. In the NCU-IISR system, Lai et al., 2016 [10] firstly detect the functions of entities through keyword matching, and secondly use the biomedical semantic role annotation method [11] to parse sentences into a predicate-argument structure (PAS), and then convert them into Subject-Verb-Object (SVO) triplets to extract causal relations between entities. Finally, entity relations and functions are combined into BEL statements. This method ignores the sentences without SVO structure, leading to low extraction recall. Liu et al., 2019 [12] propose to divide the BEL extraction task into two sub-tasks: entity relation extraction (RE) and entity function detection (FD). They use two separate BiLSTM models based on the attention mechanism to extract entity relations and detect entity functions during the BEL statement construction. Following their work, Shao et al., 2021 [13] further propose a BEL statement extraction method based on Simplified Biological Expression Language (SBEL), which acts as an intermediate form between BEL statements and entity relations/functions to retain relation and function instances as many as possible. Moreover, BERT models are applied to improve the performance of relation extraction and function detection due to their language representation superiority. However, their treatment of relation extraction and function detection as two independent sub-tasks ignores their commonality and correlations. As suggested by Liu et al., 2019 [12] that the precision of function detection dominates the contribution of the function detection to the BEL extraction, low precision in function detection deteriorates the overall performance in BEL statement extraction. Therefore, they simply use a threshold filtering approach to find more accurate entity functions to formulate the BEL statements.

3. BEL vs SBEL

A BEL statement contains Terms, Functions, and Relations. A BEL Term represents either an abundance of a biological entity or a biological process. A Function takes terms as its arguments and encodes their state information. The causal relations of *increases* or *decreases* indicate that the subject promotes or inhibits the object. Table 1 lists 6 categories of functions, their abbreviations and definitions in the corpus of the BioCreative-V shared task 4.

To simplify the task of BEL extraction, we follow the line of Shao et al., 2021 [13] to adopt SBEL statements as the bridge between BEL statements and machine learning instances. An SBEL statement expresses a causal relation between a subject and an object both with at most one function. It can be encoded in a quintuple $<$ *func1*, *e1*, *rel*, *func2*, *e2*$>$, where *e1* and *e2* are the entity mentions of the subject and the object, *func1* and *func2* are their corresponding functions and *rel* is the relation between *e1* and *e2*. Their basic idea is to convert BEL statements on the training set into SBEL statements for model training, and then conduct the SBEL statement predictions on the test set and further assemble the predicted results into BEL statements. SBEL simplifies the BEL extraction task while retaining most of the information in BEL statements, which effectively improves the BEL extraction performance. It is assumed that a causal relation only exits between a pair of entities, and each entity has at most one function. Particularly, for the function type *complex* that has more than one entity, the instance is accordingly decomposed into multiple function instances, and thus new multiple SBEL statements are formed subsequently.

Fig. 1 illustrates an example sentence (SEN: 10029842) with BEL and SBEL statements as well as the learning instances in the separate models and the joint model, respectively. There are two BEL statements (BEL1 and BEL2) among three protein entities (MGI:Akt1, MGI:Pde3b, and MGI:Ins2), e.g., the BEL statement of "p(MGI:Akt1) *increases* p(MGI:Pde3b, *pmod*(P))" means that the MGI:Akt1 protein promotes the phosphorylation of the MGI:Pde3b protein. These two BEL statements

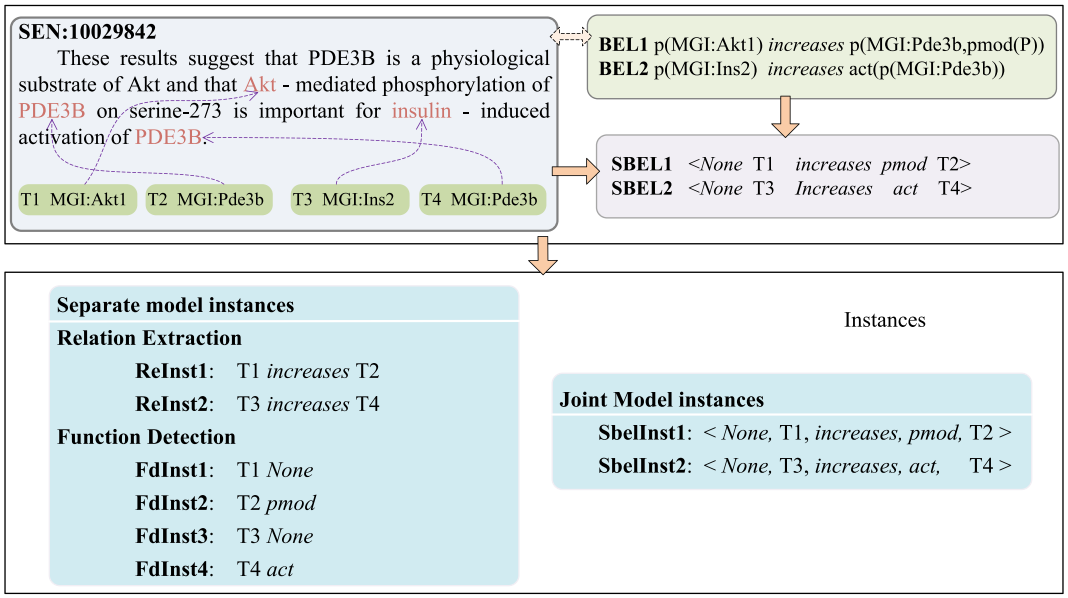


Fig. 1. The learning instances of the separate models and joint model.

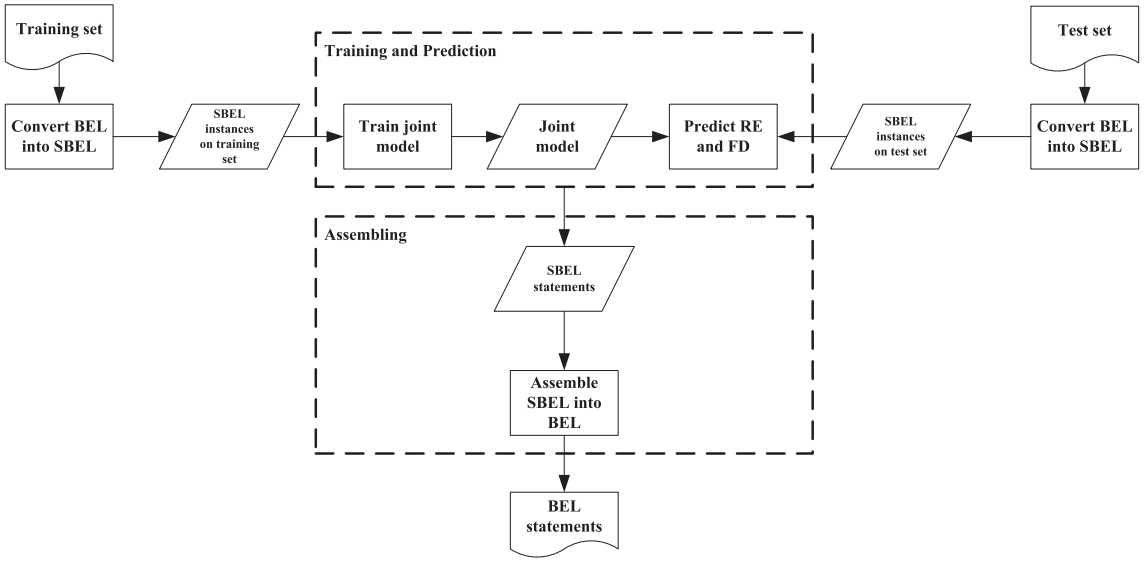


Fig. 2. Framework of joint extraction of BEL statements based on SBEL statements.

are converted into SBEL1 and SBEL2 statements between four entity mentions T1, T2, T3, and T4, respectively. For example, SBEL1 of "<*None* T1 *increases pmod* T2>" denotes that the entity mention T1(MGI: Akt1) with the *None* function *increases* the entity mention T2(MGI: Pde3b) with the *pmod* function. Note that BEL statements describe the relations among entities with identifiers while SBEL statements denote the relations between two entity mentions.

Shao et al., 2021 [13] decompose these two SBEL statements into two relation instances (ReInst1, ReInst2) and four function instances (FdInst1, FdInst2, FdInst3, and FdInst4), and use two independent models to address these two sub-tasks. The disadvantage of the separate models is that they cannot consider the inter-relationship between two sub-tasks. Different from their work [13], our approach is to address the sub-tasks of relation extraction and function detection simultaneously by training one joint model directly with SBEL instances (SbelInst1 and SbelInst2 in Fig. 1). In this way, the model can consider and exploit the inter-relationship between two sub-tasks, e.g., when the relation between two entities does not exist, their functions cannot exist.

4. Methodology

4.1. Framework

The framework of causal relation extraction based on joint learning is shown in Fig. 2. Different from the separate models (Shao et al., 2021 [13]), it is not necessary to decompose SBEL into relation instances and function instances. On the contrary, SBEL instances are directly used to train a joint-learning model. In testing, the relation and functions of a pair of entity mentions are predicted simultaneously. The main steps are as follows:

(1) **Conversion**: BEL statements are converted into corresponding SBEL instances. This step is the same as Shao et al., 2021 [13].
(2) **Training**: SBEL instances in the training set are used to train the joint learning model. The next subsection will describe this step.
(3) **Prediction**: For each pair of entities, the relation between them and the function of each entity are predicted simultaneously. If

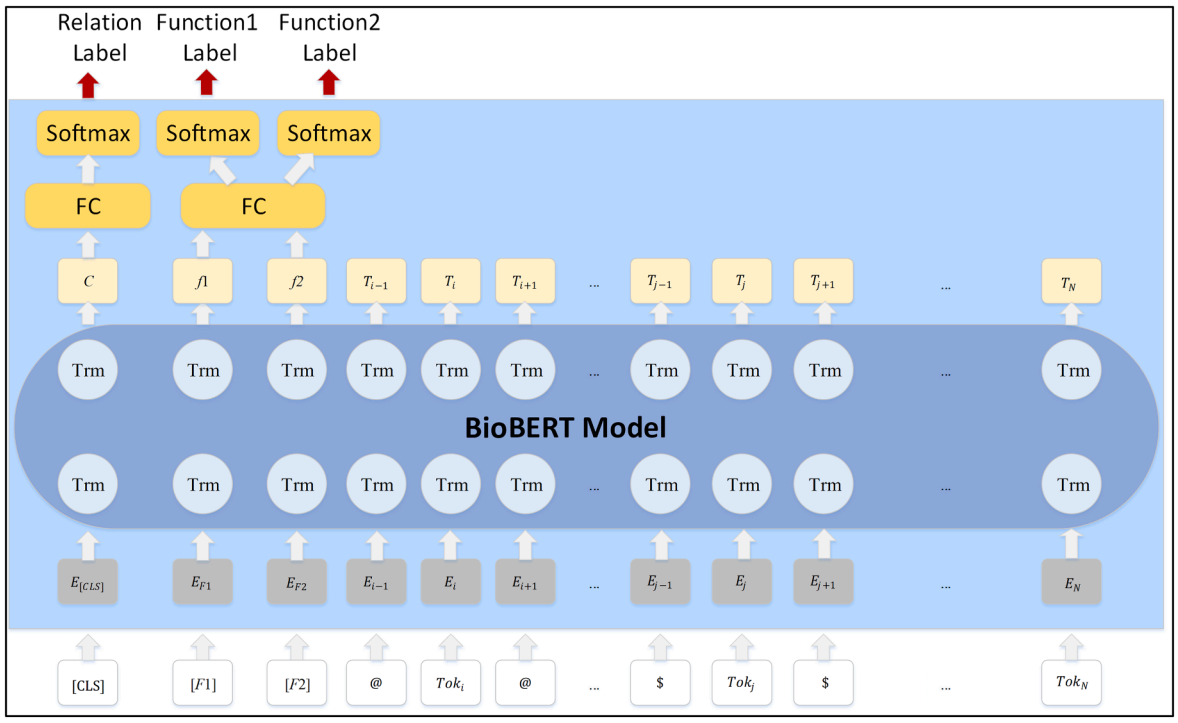


Fig. 3. Joint learning model based on BioBERT.

Table 2
BC-V BEL corpus statistics.

Statistics	Training		Test	
	SEP	**JNT**	**SEP**	**JNT**
Sentence	6,353	–	105	–
BEL	11,066	–	202	–
SBEL	10,097	–	203	–
Relations	10,097	–	203	–
increases	7,382	–	150	–
decreases	2,715	–	53	–
Functions	6,476	7,759	69	87
act	4,637	5,497	27	32
deg	119	137	6	6
pmod	712	832	5	5
sec	217	251	4	4
tloc	63	71	5	4
complex	728	971	22	36

Table 3
Five evaluation levels for BEL and SBEL statements.

Evaluation Levels		**Abbr.**	**Descriptions**
BEL	BEL Statement w/o function	BEL (Rel)	Evaluates whether the BEL statement generated only from a predicted relation triplet (without functions) is a correct BEL statement.
	BEL Statement	BEL	Evaluates whether a predicted BEL statement combining the relation and functions is correct.
SBEL	Relation	SBEL-RE	Evaluates whether the relation triplet < em1, relation, em2 > in a predicted SBEL statement is correct, ignoring its functions.
	Function	SBEL-FD	Evaluates whether the predicted functions of the subject and the object (em1 and em2) are correct when a relation holds between them.
	SBEL Statement	SBEL	Evaluates whether the whole predicted SBEL statement, i.e., the SBEL quintuple, is correct.

Table 4
Hyper-parameters of neural networks.

Hyper-parameters	**Value**
batch_size	16
epoch	3
max_seq_len	128
loss function	categorical_crossentropy
learning rate	1e-5
optimizer	Adam
development ratio	0.1
function type weight	[1,1,1,1,1,1,3]

Table 5
Performance comparison of different models in the Stage 2 evaluation.

Evaluation levels	**SEP**			**JNT**		
	P	**R**	**F1**	**P**	**R**	**F1**
SBEL-RE	**74.7**	62.0	67.8(±1.8)	74.5	**63.2**	**68.3(±1.3)**
SBEL-FD	57.7	25.8	35.6(±3.4)	**66.0**	**29.3**	**40.5(±1.5)**
SBEL	52.5	46.4	49.3(±1.4)	**57.3**	**48.7**	**52.6(±0.7)**
BEL (Rel)	59.8	46.2	52.1(±1.2)	**60.9**	**48.3**	**53.9(±1.4)**
BEL	64.1	49.6	55.9(±3.3)	**64.5**	**51.2**	**57.0(±1.1)**

Table 6
Performance comparison of different models in Stage 1 evaluation.

Evaluation levels	SEP			JNT		
	P	**R**	**F1**	**P**	**R**	**F1**
SBEL-RE	47.8	56.6	51.8(±2.0)	**49.9**	**60.1**	**54.5(±1.2)**
SBEL-FD	57.9	23.2	33.0(±6.0)	**75.1**	**31.3**	**44.1(±2.8)**
SBEL	35.3	41.8	38.3(±2.8)	**38.9**	**46.8**	**42.4(±0.7)**
BEL (Rel)	34.6	29.8	32.0(±1.2)	**35.3**	**31.2**	**33.1(±1.2)**
BEL	38.0	32.6	35.0(±2.7)	**39.6**	**35.3**	**37.3(±1.0)**

Table 7
Performance comparison with other systems.

Systems	BEL	
	Stage 1	**Stage 2**
Rule-based [5]	**39.2**	25.6
Event-based [7]	20.2	35.2
NCU-IISR [10]	19.7	33.1
Att-BiLSTM [12]	21.3	46.9
SBEL-BERT [13]	30.1	54.8
Ours (JNT)	37.3	**57.0**

Table 8
Performance comparison between SEP and JNT at the SBEL-FD level.

Function types	Percent (%)	SEP			JNT		
		P	**R**	**F1**	**P**	**R**	**F1**
act	36.8	50.2	**36.0**	**41.8**	**52.6**	33.1	40.6
complex	41.4	78.5	**35.8**	**49.2**	**100.0**	8.3	15.2
deg	6.9	**100.0**	**70.0**	**82.2**	**100.0**	60.0	74.7
pmod	5.7	6.7	5.0	5.7	**80.3**	**76.0**	**77.1**
sec	4.6	35.2	75.0	47.8	**66.4**	**85.0**	**74.0**
tloc	4.6	43.3	40.0	40.8	**60.7**	**48.0**	**52.9**
SBEL-FD	100.0	57.7	25.8	35.6	**66.0**	**29.3**	**40.5**

Table 9
Official evaluation levels for BEL statements.

Evaluation Levels	**Abbr.**	**Descriptions**
Term	–	Evaluates whether the predicted terms (entity name, type and namespace) are correct.
Function	Func	Evaluates whether the predicted terms and their associated functions are correct.
Function-Secondary	FS	Evaluates whether the predicted functions are correct, ignoring the associated terms.
Relation	Rel	Evaluates whether the relation triplet < sub, rel, obj > in a predicted BEL statement is correct, ignoring terms' functions.
Relation-Secondary	RS	Evaluates whether any two elements of the triplet < sub, rel, obj > in a predicted BEL statement are correct, ignoring terms' functions.
BEL Statement	BEL	Evaluates whether a predicted BEL statement combining the relation and functions is correct.

there exists a relation between them, the predicted SBEL statements are assembled into the corresponding BEL statements.

4.2. Joint learning of SBEL extraction

Due to the great success of the pre-trained language model in natural language processing, e.g., BERT [26], and the derived BioBERT [27] in relation extraction in biomedical text mining, we adopt the BioBERT model as the encoder for joint learning. BERT is a masked language model using the transformer [28] as a feature extractor. The self-attention mechanism enables the BERT model to have a broader global view, so it can better capture sentential semantic information.

The pre-training corpus of BERT is BooksCorpus [29] and English Wikipedia datasets. However, biomedical literature contains a large amount of domain knowledge, and the general domain pre-training language model may not perform well in biomedical information mining. Therefore, we adopt the BioBERT pre-trained model [27], which uses PubMed abstract and PMC full text as the pre-training datasets. It has achieved excellent performance in many text mining tasks in the biomedical field. Fig. 3 shows the model diagram of joint learning based on BioBERT, which combines relation extraction and function detection. We describe the input/output layers and the training loss function as follows.

Table 11
Performance comparison with other systems.

(a) Stage 1 evaluation

Systems	Term	FS	Func	RS	Rel	BEL
Rule-based[5]	62.9	55.4	**42.6**	73.3	**49.2**	**39.2**
Event-based[7]	34.0	10.0	8.6	25.1	41.4	20.2
NCU-IISR[10]	45.0	9.5	2.7	56.7	26.4	19.7
Att-BiLSTM[12]	58.6	34.3	17.7	62.3	31.6	21.3
SBEL-BERT[13]	59.8	59.6	28.5	72.2	40.4	30.1
Ours (JNT)	**68.7**	**62.6**	41.9	82.3	48.6	37.3
Ours (JNT_CW)	68.4	41.5	28.7	**82.6**	47.6	37.2

(b) Stage 2 evaluation

Systems	Term	FS	Func	RS	Rel	BEL
Rule-based[5]	82.4	56.5	30.0	82.4	65.1	25.6
Event-based[7]	54.3	26.1	20.8	61.5	43.7	35.2
NCU-IISR[10]	55.2	–	–	63.5	44.6	33.1
Att-BiLSTM[12]	**97.2**	34.8	26.6	**96.5**	65.8	46.9
SBEL-BERT[13]	94.2	**63.2**	47.9	95.8	74.3	54.8
Ours (JNT)	91.8	58.1	**50.0**	94.2	**75.3**	57.0
Ours (JNT_CW)	91.8	46.5	43.1	93.5	75.0	**58.4**

Table 10
Performance comparison of different models with and w/o function weighting.

(a) Stage 1 evaluation

Evaluation levels	SEP			SEP_CW			JNT			JNT_CW		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
SBEL-RE	47.8	56.6	51.8(±2.0)	47.8	56.6	51.8(±2.0)	**49.9**	**60.1**	**54.5**(±1.2)	48.8	58.3	53.1(±1.6)
SBEL-FD	57.9	23.2	33.0(±6.0)	62.1	19.4	29.4(±5.5)	75.1	**31.3**	**44.1**(±2.8)	**83.3**	14.5	24.7(±1.4)
SBEL	35.3	41.8	38.3(±2.8)	34.9	45.2	39.5(±1.1)	**38.9**	**46.8**	**42.4(±0.7)**	37.0	44.2	40.2(±1.2)
BEL (Rel)	34.6	29.8	32.0(±1.2)	34.8	**32.6**	**33.6(±0.8)**	**35.3**	31.2	33.1(±1.2)	34.8	30.7	32.6(±1.4)
BEL	38.0	32.6	35.0(±2.7)	37.4	35.1	36.2(±0.8)	**39.6**	**35.3**	**37.3(±1.0)**	**39.6**	35.2	37.2(±1.2)

(b) Stage 2 evaluation

Evaluation levels	**SEP**			**SEP_CW**			**JNT**			**JNT_CW**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
SBEL-RE	**74.7**	62.0	67.8(±1.8)	**74.7**	62.0	67.8(±1.8)	74.5	**63.2**	**68.3(±1.3)**	74.0	62.2	67.6(±1.7)
SBEL-FD	57.7	25.8	35.6(±3.4)	60.7	16.8	26.2(±2.7)	66.0	**29.3**	**40.5(±1.5)**	**72.6**	17.0	27.5(±1.2)
SBEL	52.5	46.4	49.3(±1.4)	55.0	45.7	49.9(±1.3)	**57.3**	**48.7**	**52.6(±0.7)**	56.9	47.9	52.0(±1.1)
BEL (Rel)	59.8	46.2	52.1(±1.2)	59.8	46.2	52.1(±1.2)	**60.9**	**48.3**	**53.9(±1.4)**	**60.9**	**48.3**	**53.9(±1.3)**[4]
BEL	64.1	49.6	55.9(±3.3)	63.3	51.0	56.5(±1.2)	64.5	51.2	57.0(±1.1)	**66.0**	**52.4**	**58.4(±1.5)**

[4]The BEL (Rel) performance scores of JNT and JNT_CW models are the same except their deviations, which is only coincidental.

4.2.1. Input layer

Given a sentence S $= \{w_1, w_2, \cdots, w_M\}$, where w_i is the i-th word and M is the number of words in the sentence. The sentence is tokenized into word pieces and three special tokens are inserted at the beginning to generate a sequence of tokens $Tokens = \{[CLS], [F1], [F2], tok_1, tok_2, \cdots, tok_N\}$ as the model input, where N denotes the number of tokens in the sentence. [CLS] is used as the classification token for relation extraction, [F1] and [F2] are used as the classification tokens for function detection of two entities respectively. Unlike [CLS], [F1] and [F2] are not inherent tokens in the BERT vocabulary, so we replace them with two reserved symbols [unused1] and [unused2] in the BERT vocabulary. And $E = \{E_{[CLS]}, E_{[F1]}, E_{[F2]}, E_1, E_2 \cdots E_N\}$ is the sequence of token embeddings. Similar to BioBERT [27], two special symbols '@' and '$' are inserted around entities to mark subject and object respectively.

4.2.2. Output layer

Each token in *Tokens* is encoded by BERT to generate the hidden representation sequence $H = \{C, f1, f2, H_i, H_{i+1} \cdots H_N\}$. $C, f1, f2$ are passed to two fully connected layers and three *softmax* classifiers to predict the relation type in the relation label set $L_r = \{l_1, l_2, \cdots, l_{n_r}\}$ and function types in the function label set $L_f = \{l_1, l_2, \cdots, l_{n_f}\}$ respectively, i.e., $[\hat{y}_r, \hat{y}_{f1}, \hat{y}_{f2}]$. They can be formulated as follows:

$$\hat{y}_r = softmax(W_r C + b_r) \tag{1}$$

$$\hat{y}_{f1} = softmax(W_f f1 + b_f) \tag{2}$$

$$\hat{y}_{f2} = softmax(W_f f2 + b_f) \tag{3}$$

Where $W_r \in \mathbb{R}^{d_z \times n_r}$, $W_f \in \mathbb{R}^{d_z \times n_f}$, d_z is the dimension of the BERT embeddings, $n_r = 3$ is the number of relation types, $n_f = 7$ is the number of function types, b_r, b_f are the bias. Note that the fully connected layer of entity 1 and entity 2 functions shares the same set of parameters $[W_f, b_f]$.

4.2.3. Loss function

The loss function of the joint model is the sum of the cross-entropy loss of relation, entity 1 function, and entity 2 function, as shown in formula (4). As the previous research work [12,13] points out, the precision of entity function detection plays an important role in the performance of BEL statement extraction. Therefore, we assign different weights to different function types in the loss of entity functions. Specifically, the more the penalty coefficient of the *None* function type, the more the improvement in the precision of entity function detection. The model training adopts the Adam optimization algorithm to optimize in the direction of loss minimization. The training loss can be formalized as below:

$$J = -\left(\sum_{i=1}^{N} \sum_{j=1}^{n_r} y_{j,r} log \hat{y}_{j,r} + \sum_{i=1}^{N} \sum_{j=1}^{n_f} w_j y_{j,f1} log \hat{y}_{j,f1} + \sum_{i=1}^{N} \sum_{j=1}^{n_f} w_j y_{j,f2} log \hat{y}_{j,f2} \right) \tag{4}$$

Here, N is the length of the input token sequence, $y_{j,r}$ is the gold relation label, $y_{j,f1}, y_{j,f2}$ are the gold function labels for entity 1 and entity 2, respectively. $w_j \in \mathbb{R}^{n_f}$ represents the weight of the j-th function type.

5. Experimentation

5.1. Datasets

The BioCreative-V BEL task dataset was selected from two corpora provided by Selventa[1] and the sbv IMPROVER Network Verification Challenge.[2] It includes one training set and one test set, where each sentence is annotated with one or more BEL statements. Table 2 reports the statistics on sentences, BEL statements, SBEL statements, and relations and functions derived from SBEL statements in our joint model (denoted as JNT) and separate models (denoted as SEP) in Shao et al., 2021[13]. The mark "-" in a cell indicates that the number in the joint model is the same as that in the separate models. It can be seen from the table that:

- The numbers of SBEL statements and relation instances are the same because both statistics are acquired based on two entities, i.e., a pair of entities only have one SBEL statement and one relation instance.
- The difference lies in the number of function instances. Generally, there are more function instances in the joint model than in the separate models because of their different statistical methods, i.e., two functions are counted in an SBEL instance while only one function is counted in a function instance.

5.2. Evaluation

We adopt commonly-used evaluation metrics to evaluate the extraction performance, namely Precision (P), Recall (R), and $F1$ measure. Due to the complexity of BEL statements, the BioCreative-V community[4] defined six official levels to evaluate how well a system performs the task of BEL extraction for various BEL elements. Among them, the BEL Statement Level is the most important since it aims to evaluate the overall performance of ultimate BEL statements. More details of these evaluation levels can be referred to [4]. Besides, we define a level of BEL Statement w/o function (BEL(Rel)) to investigate the contribution of relations alone to the performance of BEL statements.

In order to better evaluate the extraction performance of intermediate SBEL statements, we define three additional evaluation levels: SBEL-RE, SBEL-FD, and SBEL. The descriptions on these levels together with two for BEL evaluation are listed in Table 3, while others are described in Appendix. In our joint model, the test instances are SBEL statements, so the performance at three SBEL levels can be calculated directly and be helpful to examine how well these elements are extracted. In the separate models, the relations between entities and their functions are first assembled into SBEL statements, then SBEL performance can be evaluated on these SBEL statements.

The evaluation of the shared task on the test set is conducted in two stages: Stage 1 and Stage 2. In Stage 1, gold entities are not provided, so entity mentions such as genes, chemicals and diseases are automatically recognized and linked to the corresponding databases. We follow the same approach as Liu et al., 2019[12] and Shao et al., 2021[13] to name entity recognition and alignment, i.e., using GNormplus [30] for genes and proteins, tmChem [31] for chemicals, and DNorm [32] for diseases. In Stage 2, since gold entities, but not their mentions, are given, we still need to align the entities with their mentions in the text and further find unaligned entities via dictionary search, particularly for biological processes.

5.3. Models for comparison

Four different models as follows are compared to evaluate the effectiveness of joint learning and function weighting:

- **SEP**: we rerun the separate models of Shao et al., 2021[13] as a baseline. The sub-tasks of relation extraction and function detection are trained and applied to prediction separately, and then their results are assembled into SBEL statements and finally BEL statements.

[1] https://www.flagshippioneering.com/companies/selventa/.

[2] https://bionet.sbvimprover.com/.

- **JNT**: our joint model that can simultaneously train relation extraction and function detection. The SBEL statements predicted by the model are directly assembled into BEL statements.
- **SEP_CW**[3]: similar to SEP except that function weighting is applied to the loss function of the function detection model.
- **JNT_CW**: similar to JNT except that function weighting is applied to the joint loss function.

5.4. Settings

The "biobert-pubmed-v1.1" version of BioBERT is used as the pre-trained model. The hyper-parameters of the joint learning model are similar to those of separate models in Shao et al., 2021 [13] as shown in Table 4. However, the value of function weights is set to [1,1,1,1,1,1,3] for 6 plus *None* function types. The weight of the *None* type is manually-tuned on the development set from the range of 2 to 5 while other weights are fixed to 1.

5.5. Experimental results

5.5.1. Performance comparison of the SEP and JNT models in the Stage 2 evaluation

Table 5 compares the performance of SEP and JNT models on the test set in Stage 2 evaluation, where five evaluation levels of SBEL-RE, SBEL-FD, SBEL, BEL(Rel), and BEL are considered. We take the average performance of five random runs as the overall performance. The highest P/R/F1 values in each row are shown in bold, and the values in the parentheses right to the F1 measures represent the standard deviations across five runs. As can be seen from the table:

- The performance of JNT is better than that of SEP at almost all levels except the precision at SBEL-RE. Among them, the F1 score at the BEL level has been improved by 1.1 units. This demonstrates the effectiveness of joint learning over separate models.
- As for SBEL-FD, compared with SEP, the precision and recall of JNT increase by 8.3 units and 3.5 units, respectively, suggesting that JNT can effectively improve the precision of function detection without reducing its recall.

5.5.2. Performance comparison of the SEP and JNT models in Stage 1 evaluation

Table 6 compares the performance scores of SEP and JNT on the test set in Stage 1 evaluation. The experimental settings are the same as Stage 2 except that no gold entities are given and therefore the same approach to entity recognition and alignment as [12 13] is adopted. As can be seen from the table:

- Compared with the Stage 2 performance, the performance scores of all four models decrease drastically, which is obviously caused by the noise in automatically recognized named entities.
- The performance scores of two joint models are generally better than those of two separate models, and this improvement is due to that the joint model can greatly improve the precision of SBEL-FD.

5.5.3. Performance comparison with other systems

Table 7 compares the BEL performance of our joint model with other systems on the BC-V BEL extraction task both in Stage 1 and Stage 2 evaluation, which are rule-based [5], event-based [7], and other machine learning methods i.e., NCU-IISR [10], Att-BiLSTM [12], and SBEL-BERT [13]. The highest performance scores in Stage 1 and Stage 2 respectively are shown in bold. As can be seen from the table:

- In Stage 1, our JNT model still achieves promising performance at the BEL level, which is close to the rule-based [5] and significantly outperforms other systems. This may be due to that the joint model takes into account both relation extraction and function detection, and thus has better noise tolerance to entities.
- In Stage 2, our JNT model achieved the highest performance at the BEL level. The F1 score of our JNT is 57.0 %, at least 2.2 units higher than those of other systems. Since the BEL performance is the overall evaluation metric, this demonstrates the effectiveness of the joint learning method in BEL statement extraction.

6. Discussion and analysis

This section compares and analyzes Stage 2 performance on the test set from the perspectives of joint learning, and further discusses the possible reasons behind the performance improvements.

It can be seen from Table 5 that the performance scores at SBEL-RE between the separate models and the joint models are similar. Therefore, in Table 8, we only compare the performance differences in function detection at the SBEL-FD level in Stage 2 evaluation. Similarly, the higher P/R/F1 scores in each row among the two models are shown in bold.

It is observed from the Table 8 that the reason why the joint model is superior to the separate models in overall P/R/F1 scores is that the former has higher precision scores in all types than the latter. Although the recall scores of *act*, *complex*, and *deg* function types decrease, those of the other three types are significantly greater than the latter.

By observing and comparing the function instances predicted by the two models, we find that for function types of *pmod*, *sec* and *tloc* with relatively fewer instances, the separate model cannot predict correctly as the joint model can do. The reason may be that the function instance generation on the training set for the separate model ignores possible multiple functions of the same entities, resulting in a large number of entities with function clue words not correctly annotated. This interferes with the training of the function detection model and affects the generalization ability of the model. In the joint model, relation extraction and function detection are jointly trained, and those entity functions derived from the relation label of *None* will not interfere with the training phase, and the generalization ability of the joint model is thus significantly improved.

Taking the *pmod* function type as an example, among the training instances of the separate model for function detection, nearly 50 % of the entities with the keyword "phosphorylation" are not labeled as *pmod* function type because they do not participate in any relation. For example, in the sentence "These effects (TLR2,3,5,6 activation of VEGF and IL8) were prevented by treatment with a selective inhibitor of ***EGFR*** phosphorylation (AG-1478), a metalloprotease (MP) inhibitor, a reactive oxygen species (ROS) scavenger, and an NADPH oxidase inhibitor." (PMID:18375743), the protein entity ***EGFR*** that should have the *pmod* function does not involve any relation instances with other entities.

7. Conclusion and future work

This paper proposes a joint learning model of relation extraction and function detection based on SBEL statements, which aims to capture the potential relationship between two sub-tasks. Function type weighting is also applied to the joint loss function, so as to improve the precision of function detection and further improve the performance of BEL causal relation extraction. Experimental results show that the joint model mitigates the issue caused by the independence of relation extraction and function detection in the separate models, that is, when the relation between two entities does not exist, the function types of the two entities are forcibly marked as *None*, regardless of the true function types of the entities in the sentence. Because the noise in the training instances of function detection is eliminated, and the interaction between relation and function is fully utilized, joint learning significantly improves the

[3] The performance results of SEP_CW and JNT_CW are reported in Table 10 in Appendix.

precision and recall of function detection, and thus improves the performance of BEL statements.

Similar to the previous BEL statement extraction work, although the performance at the function level is significantly improved, it is still far from satisfactory. One limitation of our work is that entity names, which play an important role in function detection, are masked in the input text. Future work will consider incorporating entity names and expanding entity acronyms to their full names in order to improve function detection performance. Due to the success of seq2seq models in information extraction, particularly in relation extraction [33], and the nature of multiple mentions for the same entity in BEL statements, another future work can adopt seq2seq models to recast BEL extraction as direct generation of BEL statements from a sentence.

Funding

This research is supported by the National Natural Science Foundation of China [61976147; 2017YFB1002101] and the research grant of The Hong Kong Polytechnic University Projects [#1-W182].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Official evaluation levels for BEL extraction

Table 9 explains all official evaluation levels for BEL extraction used in our models, including Term Level (Term), Function Level (Func), Function Secondary Level (FS), Relation Level (Rel), Relation Secondary Level (RS), and BEL Statement Level (BEL).

Performance comparison between models with and without function weighting

Table 10 (a) and (b) compare the performance of SEP, SEP_CW, JNT, and JNT_CW models on the test set in the Stage 1 and Stage 2 evaluations respectively, where five evaluation levels of SBEL-RE, SBEL-FD, SBEL, BEL(Rel), and BEL are considered. Similarly, the highest P/R/F1 values in each row are shown in bold, and the values in the parentheses right to the F1 measures represent the standard deviations across five runs. The table shows that:

In the Stage 1 evaluation, whether for the joint models or the separate models, function weighting consistently improves the precision of SBEL-FD. However, in separate models, it improves the overall performance of BEL statements, while in the joint model, it does not. This is probably due to the noise introduced in the Stage 1 evaluation, function weighting decreases the performance of relation extraction while increasing the precision of function detection, the BEL(Rel) and BEL performance scores decrease accordingly.

In the Stage 2 evaluation, when function weighting is applied in SEP_CW and JNT_CW, compared to SEP and JNT, the precision scores at SBEL-FD are improved by 3 and 6.6 units respectively. Although the recall and F1 scores of SBEL-FD decrease, the F1 scores at BEL of SEP_CW and JNT_CW are still better than those of SEP and JNT without function weighting. This is because the function weighting significantly reduces the number of wrongly recognized entity functions in SBEL statements, while there is not any significant difference at SBEL-RE among these four models.

Performance comparison of other systems at the official evaluation levels

Table 11 (a) and (b) compare the performance with other systems at the official evaluation levels in the Stage 1 and Stage 2 evaluations respectively. The best performance scores at each level are showed in boldface. It shows that in Stage 1 our scores at all levels are comparable with the best rule-based one [5], while in Stage 2 our system outperforms others at the Func, Rel, and BEL levels.

References

[1] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, J. Wang, The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models, Bioinformatics 19 (4) (2003) 524–531.
[2] E. Demir, M.P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G.D. Bader, The BioPAX community standard for pathway data sharing, Nature biotechnology 28 (9) (2010) 935–942.
[3] T. Slater, D. Song, Saved by the BEL: ringing in a common language for the life sciences, Drug Discovery World Fall 2012 (2012) 75–80.
[4] F. Rinaldi, T.R. Ellendorff, S. Madan, S. Clematide, A. Van der Lek, T. Mevissen, J. Fluck, BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language, Database 2016 (2016) 1–15.
[5] K.E. Ravikumar, M. Rastegar-Mojarad, H. Liu, BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences, Database 2017 (2017) 1–12.
[6] Ravikumar, K.E., Wagholikar, K.B., & Liu, H. (2014). Towards pathway curation through literature mining–a case study using PharmGKB. In Pacific Symposium on Biocomputing 2014 (pp. 352-363).
[7] M. Choi, H. Liu, W. Baumgartner, J. Zobel, K. Verspoor, Coreference resolution improves extraction of Biological Expression Language statements from texts, Database 2016 (2016) 1–14.
[8] Björne, J., & Salakoski, T. (2011). Generalizing Biomedical Event Extraction. In Proceedings of BioNLP Shared Task 2011 Workshop, pages 183–191, Portland, Oregon, USA. Association for Computational Linguistics.
[9] Fluck, J., Klenner, A., Madan, S., Ansari, S., Bobic, T., & Hoeng, J. (2013). BEL networks derived from qualitative translations of BioNLP Shared Task annotations. In Proceedings of the 2013 Workshop on Biomedical Natural Language Processing, pages 80–88, Sofia, Bulgaria. Association for Computational Linguistics.
[10] P.T. Lai, Y.Y. Lo, M.S. Huang, Y.C. Hsiao, R.T.H. Tsai, BelSmile: a biomedical semantic role labeling approach for extracting biological expression language from text, Database 2016 (2016) 1–9.
[11] R.T.H. Tsai, P.T. Lai, A resource-saving collective approach to biomedical semantic role labeling, BMC bioinformatics 15 (1) (2014) 1–12.
[12] S. Liu, W. Cheng, L. Qian, G. Zhou, Combining relation extraction with function detection for BEL statement extraction, Database 2019 (2019) 1–12.
[13] Y. Shao, H. Li, J. Gu, L. Qian, G. Zhou, Extraction of causal relations based on SBEL and BERT model, Database 2021 (2021) 1–12.
[14] J. Baxter, A Bayesian/information theoretic model of learning to learn via multiple task sampling, Machine learning 28 (1) (1997) 7–39.
[15] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).
[16] Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-Task Deep Neural Networks for Natural Language Understanding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4487-4496).
[17] Søgaard, A., & Goldberg, Y. (2016). Deep multi-task learning with low level tasks supervised at lower layers. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 231-235).
[18] Sanh, V., Wolf, T., & Ruder, S. (2019). A hierarchical multi-task approach for learning embeddings from semantic tasks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6949-6956).
[19] Duong, L., Cohn, T., Bird, S., & Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers) (pp. 845-850).
[20] Yang, Y., & Hospedales, T. (2017). Trace Norm Regularised Deep Multi-Task Learning. In 5th International Conference on Learning Representations, Toulon, France.
[21] M. Miwa, M. Bansal, End-to-end relation extraction using LSTMs on sequences and tree structures, In Association for Computational Linguistics (ACL) (2016) 1105–1116.
[22] G. Bekoulis, J. Deleu, T. Demeester, C. Develder, Adversarial training for multi-context joint entity and relation extraction, In Empirical Methods in Natural Language Processing (EMNLP) (2018) 2830–2836.
[23] Luan, Yi., Wadden, D., He, L., Shah, A. Ostendorf, M., & Hajishirzi, H. (2019). A general framework for information extraction using dynamic span graphs. In North American Chapter of the Association for Computational Linguistics (NAACL), pages 3036–3046.
[24] Wadden, D., Wennberg U., Luan, Y., & Hajishirzi, H. (2019). Entity, Relation, and Event Extraction with Contextualized Span Representations. 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), pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

[25] Lin, Y., Ji, H., Huang, F., & Wu, L. (2020). A Joint Neural Model for Information Extraction with Global Features. meeting of the association for computational linguistics.
[26] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
[27] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pages 6000-6010, 2017.
[29] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In 2015 IEEE International Conference on Computer Vision (ICCV), 2015 (pp. 19-27).
[30] C. Wei, H. Kao, Z. Lu, GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains[J], Biomed Research International 2015 (2015) (2015) 1–7.
[31] Leaman, R., Wei, C., & Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization[J]. Journal of Cheminformatics. 2015, 7(1 supplement): 1-10.
[32] Leaman, R., Islamaj, D., & Lu, Z. (2013). DNorm: disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22):2909-2917.
[33] Giorgi, J., Bader, G., & Wang, B. (2022). A sequence-to-sequence approach for document-level relation extraction. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 10–25, Dublin, Ireland. Association for Computational Linguistics.