How Much do Knowledge Graphs Impact Transformer Models for Extracting Biomedical Events?

Laura Zanella and **Yannick Toussaint**
LORIA (Université de Lorraine, CNRS, Inria)
54506 Vandoeuvre-lès-Nancy, France
{laura-alejandra.zanella-calzada, yannick.toussaint}@loria.com

Abstract

Biomedical event extraction can be divided into three main subtasks; (1) biomedical event trigger detection, (2) biomedical argument identification and (3) event construction. This work focuses in the two first subtasks. For the first subtask we analyze a set of transformer language models that are commonly used in the biomedical domain to evaluate and compare their capacity for event trigger detection. We fine-tune the models using seven manually annotated corpora to assess their performance in different biomedical subdomains. SciBERT emerged as the highest performing model, presenting a slight improvement compared to baseline models. Then, for the second subtask we construct a knowledge graph (KG) from the biomedical corpora and integrate its KG embeddings to SciBERT to enrich its semantic information. We demonstrate that adding the KG embeddings to the model improves the argument identification performance by around 20 %, and by around 15 % compared to two baseline models. Our results suggest that fine-tuning a transformer model that is pretrained from scratch with biomedical and general data allows to detect event triggers and identify arguments covering different biomedical subdomains, and therefore improving its generalization. Furthermore, the integration of KG embeddings into the model can significantly improve the performance of biomedical event argument identification, outperforming the results of baseline models.

1 Introduction

Biomedical event extraction is a complex information extraction task that identifies key information from large sets of textual data for further applications, such as the study of biomolecular mechanisms or epigenetic changes. A biomedical event is constructed from an event trigger and one or more arguments that orbit around the trigger. Event triggers generally refer to nouns or verbs that express an action, circumstance or eventuality, while the arguments refer either to biomedical entities or to other events, called nested events. Figure 1 [1] shows the example of a sentence containing two biomedical events, '-Reg' (which stands for 'Negative regulation') and 'Locl' (which stands for 'Localization'). The event constructed from the trigger word 'excretion' of type 'Locl' (the event is given the same type as the trigger) presents as single argument the biomedical entity of type 'D/C' (which stands for 'Drug or compound'), playing the role of 'Th' (which stands for 'Theme'). This role allows answering the question 'What is excreted?'. On the other hand, the event constructed from the trigger word 'reduces' of the type '-Reg', presents two arguments. The first argument is a biomedical entity of the type 'Drug or compound', playing the role of 'Cause'. This role allows answering the question 'What causes the reduction?'. The second argument is the nested event 'Locl' described before, playing the role of 'Theme', answering the question 'What is reduced?'.

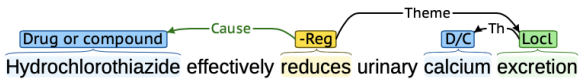

Figure 1: Example of event extraction; the '-Reg' ('negative regulation) event has the 'Locl' (localization) nested event as argument.

Event extraction is usually divided into three main subtasks, event trigger detection, argument identification and event construction. Event trigger detection identifies and classifies the trigger words into a set of predefined types of event triggers, while argument identification identifies and classifies the roles between the event triggers and their respective arguments (Shen et al., 2019). Event construction refers to the unmerging of the arguments that correspond to the same event for its

[1] The visualization of the annotated sentence is done using the visualization tool *brat* https://brat.nlplab.org/.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 145–155

construction (Björne and Salakoski, 2011).

Event trigger detection has a fundamental role in the construction of events. Indeed, the triggers are the targets that allow us to know that an event may exist (Cui et al., 2020). This subtask is usually considered as a classification problem, where each word need to be classified into a predefined set of trigger types. Difficulty for trigger detection comes from the sensitivity to the domain or subdomain (text can present specialized language), linguistic forms (triggers can be single words, multi-words, discontinuous markers) and ambiguity on the trigger class (a trigger word can be given different trigger classes) (Zerva and Ananiadou, 2015).

Argument identification can be also considered as a multi-category classification problem, where the directed relation between a trigger and an entity or other event needs to be classified into a predefined set of role types. When these arguments are correctly identified, the event extracted has the potential to provide a reliable means of improving domain knowledge. One of the main complexities in identifying arguments is that they can be part of one or multiple events (one-to-one and one-to-multiple relations), where they play the same or different roles.

Following (Ramponi et al., 2020), event trigger detection is the main source of errors in event extraction, where around 31 % of the errors correspond to non-detection of triggers and 28 % to over-detection of triggers. Further, the non-detection of arguments represents around 23 % of errors and the over-detection of arguments around 7 %. Transformer language models have been widely adopted to try to reduce errors in event extraction due to their positive achievements in performance for solving different types of Natural Language Processing tasks. BERT (Devlin et al., 2018), which stands for Bidirectional Encoder Representations from Transformers, is a language model designed to pretrain bidirectional representations of words, taking into account the semantics by considering both left and right directions of the text. From this pretraining, BERT can be fine-tuned by including additional layers on top of the model to solve new specific tasks. Furthermore, a number of domain-specific BERT variants have been developed by being trained on large corpora with the same context, such as the biomedical domain. However, since the learning of the models is limited to the subdomain in which they were trained, they present limitations in performance when using them in different biomedical subdomains

To improve the integration of domain knowledge, knowledge graph (KG) models have been implemented along with language models for different information tasks in the biomedical domain (Huang et al., 2020; Yang et al., 2020; Dasgupta et al., 2021; Roy and Pan, 2021; Milošević and Thielemann, 2023). Biomedical KGs are a resource of integration of one or more sources of information (often manually curated datasets) into a graph, where biomedical entities can be represented by nodes and the relations between them by edges (Nicholson and Greene, 2020). KG models integrate nodes and edges into a low-dimensional vector space, known as embeddings, preserving the semantic information of the KG.

In this work, we first analyze the performance of five previously trained transformer language models to identify whether they allow the identification of triggers in different biomedical subdomains. Then, we enrich the semantic information of the best-performing model with KG embeddings to assess whether integrating these embeddings improves the model's ability to identify biomedical arguments and their roles. For this purpose, BERT, BioBERT, SciBERT, PubMedBERT, and BioMedRoBERTa are fine-tuned using two different classifiers, a linear layer and a Bidirectional Long Short Term Memory (Bi-LSTM) layer, to detect biomedical event triggers. These BERT variants are chosen for comparison since they share the same BERT architecture but have previously been pretrained with different data in the biomedical and/or general domain, showing positive results in biomedical information extraction tasks (Lee et al., 2020; Beltagy et al., 2019; Erdengasileng et al., 2022). Models are learned using seven manually annotated data sets merged together. These corpora were originally developed for the event extraction task in different biomedical subdomains. Then, a KG is constructed from the biomedical events contained in the biomedical corpora and its KG embeddings are computed. These embeddings are integrated into the transformer language model to classify the roles between the previously identified triggers and the biomedical entities and/or other triggers, in order to detect the event arguments.

1.1 Contributions

Our main contributions include (1) evaluation and comparison of five transformer language models based on BERT for the detection of biomedical event triggers, (2) proposal of a novel strategy to integrate KG embeddings into transformer language models to identify biomedical event arguments, (3) empirical analysis of the effectiveness of merging annotated corpora to detect biomedical event triggers and identify arguments on different biomedical subdomains.

2 Related Work

(Rahul et al., 2017) use Recurrent Neural Networks (RNN) to extract higher level features through the hidden state of the network to identify biomedical event triggers. They also use the word and the entity type embeddings as features, demonstrating positive results in the MLEE (Pyysalo et al., 2012) corpus. (Duan et al., 2017) and (Zhao et al., 2018) explore an augmentation of the semantic information by integrating the full document representation. Both propose the use of RNNs to extract cross-sentence features without the use of external resources. (Nguyen and Grishman, 2018) present a Graph Convolution Network (GCN) model to exploit syntactic dependency relations. They use dependency trees to link words to their informative context for event trigger detection. (Yan et al., 2019) also propose a GCN model, integrating aggregative attention to model and aggregate multi-order syntactic representations of the sentences, while in the case of (Cui et al., 2020), they extend the GCN by adding the relation aware concept, which exploits the syntactic relation labels and models the relation between words. DeepEventMine (Trieu et al., 2020) is an end-to-end system for event extraction that consists in four main modules that identify the event triggers and the arguments. For each of the modules, BERT is used as base model and a linear layer is added. It has achieved the SOTA performance on seven biomedical nested event extraction tasks. (Portelli et al., 2021) compare BERT and five of its variants for the identification of Adverse Drugs and Events (ADEs). They show that span-based pretraining from spanBERT provides an improvement in the recognition of ADEs. Besides, the pretraining of the models in the specific domain is useful in comparison to train the models from scratch. (Ramponi et al., 2020) developed BEESL, a neural network model based on a sequence labeling system for the extraction of events. The system converts the event structures into a format of sequence labeling, and uses BERT as language model. (Chen, 2021) propose the Multi-Source Transfer Learning-based Trigger Recognizer system, which is an extension on transfer learning using multiple source domains. Datasets from different domains are used for jointly train the neural network, achieving a higher recognition performance on the biomedical domain, having a wide coverage of events.

KG models have recently been also used for information extraction tasks in the biomedical domain. (Sastre et al., 2020) proposes a model based on a Bi-LSTM to extract drug information from drug labels and integrate it into knowledge graph-based embedding space to evaluate drug label accuracy. In (Huang et al., 2020) is proposed to detect relations between entities in biomedical events using a question answering approach. They incorporate domain knowledge into a pretrained language model using Graph Edge-Conditioned Attention Networks (GEANet), showing improved capabilities in inferring complex events. (Lai et al., 2021) presents a GCN network with attention for biomedical entity and relation extraction based on knowledge graphs embeddings. They first construct a KG by predicting the links between the biomedical entities and then make the predictions by merging the word entities and the embeddings. (Fei et al., 2021) proposes BioKGLM, a system where a pretrained language model is enriched by integrating large biomedical knowledge graphs. To effectively encode knowledge, they explore different fusion strategies to facilitate knowledge injection.

According to these works, transformer architectures have achieved competitive performance for extracting biomedical events, and the use of pretrained language models has shown an improvement in the performance of this task. However, non-detection, over-detection and misclassification of triggers continues being the most important cause of errors in event extraction (Ramponi et al., 2020). Besides, most of these works have been developed in a specific biomedical subdomain, not allowing a generalization to different subdomains. This is a limitation in the extraction of biomedical events because the biomedical language in texts is usually specialized and very specific.

We present an alternative approach to overpass this limitation, combining corpora from different

biomedical subdomains to train transformer language models in a broader biomedical domain. Besides, we enrich the context of the domain-specific language model SciBERT using KG embeddings. This strategy adds semantic knowledge about complex links that allow inferences between biomedical concepts that are not directly related. Our proposal outperforms significantly two strong baselines identifying arguments in the biomedical Cancer Genetics corpora.

3 Method

This work proposes the approach shown in Figure 2. The annotated data is given as input to the pretrained transformer language models for fine-tuning and then passed to a classification layer for event trigger detection. At this stage, a Named Entity Recognition (NER) task is performed to identify and classify the triggers in the text. Simultaneously, a KG is built from the annotated data and the KG embeddings are calculated. These embeddings are later integrated to the transformer language model and the fine-tuning is done to identify arguments. Here, a Relation Extraction (RE) task is performed to identify and classify the roles between the triggers and candidate arguments.

3.1 Transformer Language Models: BERT

BERT (Devlin et al., 2018) is the first and the basis of transformer language models. It is a contextualized word representation model based on a masked language model pretrained with bidirectional transformers. In BERT, the sequence of input tokens (words or sub-words) is constituted with initial vectors that are the combination of the token embeddings, the (token) position embeddings and the segment embeddings (text segment to which the token corresponds) through element-wise summation. The embeddings are then passed to a set of layers of transformer modules. Each transformer layer generates a contextual representation of every token by summing the non-linear transformation of the tokens' representations from the previous layer. This representation is weighted by the attentions calculated using the representations of the previous layer as query. The last layer generates the contextual representations for all the tokens, where the information of the whole text span is combined.

Following the BERT principle, other transformer models have been pretrained with data from specific domains, e.g. biomedical data, presenting better adaptation for solving in-domain tasks. BioBERT (Lee et al., 2020) and BioMedRoBERTa (Gururangan et al., 2020) are some examples of BERT variants pretrained in the biomedical domain.

3.2 Biomedical Trigger Detection using Transformer Language Models

Various downstream text mining tasks can be performed by making minimal modifications to the BERT architecture, through a process of fine-tuning. Here, the transformer models are fine-tuned for NER, which aims to recognize domain-specific nouns in a corpus by giving each word in a sentence a predefined class. Since here NER is adapted to detect triggers, it implies not only identifying nouns, but also verbs and in some cases adjectives. For this purpose, after obtaining the contextual representation of the tokens in vectors, a classification layer is added to classify these vectors into the event trigger classes. Two different classification layers are used separately for comparison, a linear layer and a Bi-LSTM layer. The output labels are obtained following the IOB (Inside-Outside-Beginning) tagging to classify the triggers into the predefined trigger categories (in the case of the I and B tags).

3.3 Biomedical Argument Identification using Transformer Language Models and KG Embeddings

Biomedical event argument identification refers to finding the arguments that belong to an event and the role they play in it. A strategy to identify the arguments is RE, where the goal is to capture and classify the relations between triggers and biomedical entities and/or other triggers. Here, RE is applied by enriching with the KG embeddings of the triggers, the roles and the arguments the semantic information of transformer language model.

For this, a KG is first constructed from the gold data, where the semantic types of the triggers and biomedical entities are considered as the nodes, and relations between them are considered as the directed edges from the triggers to the arguments. These relations represent the role that the argument is playing with respect to the trigger. A KG is a knowledge base presented in a graphical structure format, composed by multiple types of relations between entities. KGs can be represented as a set of facts in the form of triples (h, r, t), where h, r and t represent the head, relation and tail, respectively. In

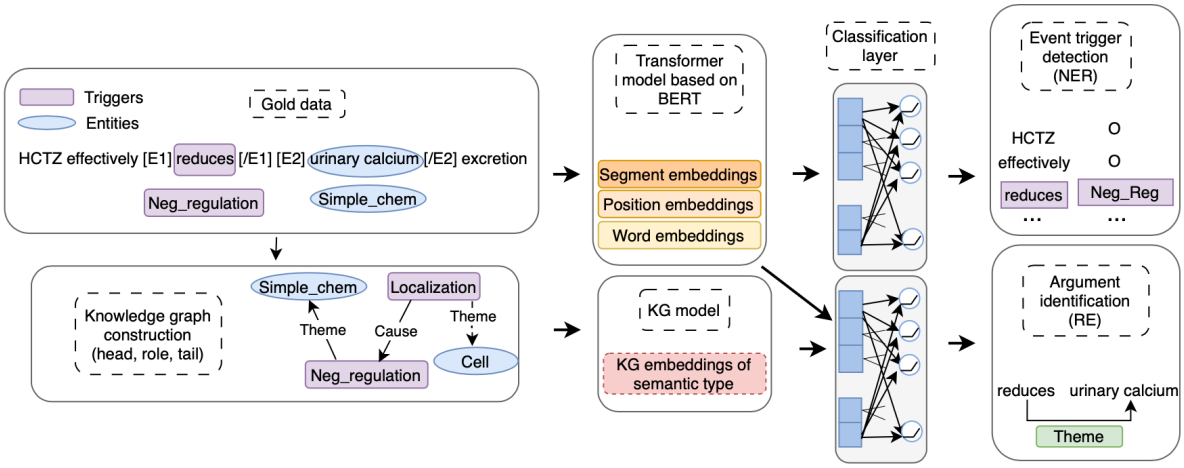

Figure 2: Overview of the approach proposed to detect event triggers and identify arguments.

order to model and infer the relations in the triples, KG embeddings are calculated, mapping the content of entities and relations to low-dimensional continuous vectors to be later used for predicting new relations (Wang et al., 2021). A scoring function is used to measure the reliability of the triples based on the embeddings, where a higher score means that the triple is more likely to be true.

KG models are evaluated by ranking all the triples according to their scores and calculating the standard evaluation metrics Mean Reciprocal Rank (MRR) and Hits@z, where $z \in 1, 3, 10$ (Kristiadi et al., 2019; Islam et al., 2021). Both metrics scores range [0,1], where the higher value demonstrates the better ranking of positive test triples, which means a better prediction performance. Here, the KG embeddings were calculated using SimplE, a model based on the tensor factorization approach, canonical polyadic decomposition (Sorber et al., 2013). It learns independent embedding vectors for the entities in the head and the tail, even if they are tied. SimplE encodes the embeddings of the two entities h and t into $\vec{h}$ and $\vec{t}$, respectively, by parameter sharing allowing to integrate the dependence between them into a relation vector $\vec{v}_r$ and $\vec{v}_{r-1}$ for the inverse relation. These embeddings are optimized by satisfying the scoring function of Equation 1.

$$\frac{1}{2}(\langle \vec{h}, \vec{v}_r, \vec{t} \rangle + \langle \vec{t}, \vec{v}_{r-1}, \vec{h} \rangle) \quad (1)$$

Finally, the KG embeddings are integrated to the embeddings of the transformer language model and then passed to a linear classification layer to identify the roles.

3.4 Incorporating KG Embeddings to Transformer Language Models

To obtain the KG embeddings, the total triples represented as *(trigger type, role, argument type)*, are randomly split into 80 % to train and 20 % to test the KG model and obtain the embeddings. These embeddings are incorporated to the contextual representation of the transformer language model following the three different strategies described below. v_{tr} and v_{ar} represent the embeddings of the trigger and the argument from the transformer language model, respectively, and kg_{tr}, kg_{ar}, kg_r represent the KG embeddings of the trigger, argument and role, respectively.

$$KG_{tr,ar} = [v_{tr}; kg_{tr}; v_{ar}; kg_{ar}] \quad (2)$$

$$KG_r = [v_{tr}; v_{ar}; kg_r] \quad (3)$$

$$KG_{tr,r,ar} = [v_{tr}; kg_{tr}; v_{ar}; kg_{ar}; kg_r] \quad (4)$$

4 Experimental Settings

4.1 Corpora

Table 1 presents the seven publicly available datasets used for fine-tuning the transformer models. These data were manually or semi-manually annotated by experts and released to be used in the development and improvement of event extraction models.

Cancer Genetics (CG) 2013 (Nédellec et al., 2013) contains information of bio-processes in the development and progression of cancer. Epigenetics and Post-translational Modifications (EPI) 2011 (Ohta et al., 2011) focuses on proteins and DNA modifications. GENIA 2011 (Kim et al., 2011) and

Table 1: Statistics of the corpora used.

Dataset	Ent types	Trig Types	Role Types	No. Events	Documents	Train/Dev/Test
CG 2013	18	40	9	17,248	PubMed abstracts	300/100/200
EPI 2011	2	15	5	2,453	PubMed abstracts	600/200/400
GENIA 2011	2	9	6	13,560	MEDLINE abstracts	908/259/347
GENIA 2013	3	13	6	6,016	PMC full-text	222/249/305
ID 2011	6	10	6	2,779	PMC full-text	152/46/118
PC 2013	4	24	8	8,121	PubMed abstracts	260/90/175
MLEE	16	26	9	6,677	PubMed abstracts	131/44/87

GENIA 2013 (Kim et al., 2013) present both information about the transcription factors in blood cells, but this last updated with more recent articles. Infectious Diseases (ID) 2011 (Pyysalo et al., 2011) consists of data about biomolecular mechanisms of infectious diseases, virulence and resistance. Pathway Curation (PC) 2013 (Nédellec et al., 2013) focuses on targets reactions relevant to the development of biomolecular pathway models. Multi-Level Event Extraction (MLEE) (Pyysalo et al., 2012) presents different levels of biological organization ranging from the subcellular to the organism level.

For the development of the experiments, the training and development datasets of all the corpora are initially merged into one single dataset and split into sentences, obtaining a total of 24,819 sentences. The sentences are then split (following the split of the KG training and test sets) to have a set of 80 % to train and 20 % to test, containing 19,855 and 4,964 sentences, respectively. We do not use the original test datasets for the experiments since the annotations are not released. All the trigger types from each dataset are considered for the final trigger classification, presenting a final set of 58 trigger types (some types are overlap among the different corpora).

4.2 Knowledge Graph Construction

The KG is constructed from the events contained at the document level from the total dataset. The nodes represent the biomedical entities and the triggers, and the edges, which are directed from the triggers to the arguments, represent the roles. Nodes contain the information about the semantic type of the biomedical entities or the triggers, while edges contain the information about the role type (e.g. (*Locl*, *Th*, *D/C*), from the event ***Locl*** in Fig. 1). All the types of triggers, biomedical entities and roles from each dataset are considered for the KG construction, presenting a final set of 81 types of triggers and biomedical entities and 12 types of roles (some types overlap among the different corpora).

From this step, 3,387 graphs are constructed. Then, the graphs are post-processed, keeping only those that present at least two nodes and one edge, reducing the total number of graphs to 2,721. All KGs are finally merged into one single graph through a disjoint union, containing 99,251 nodes and 56,931 edges.

4.3 Transformer Language Models

We compare BERT (Devlin et al., 2018), and four BERT variants pretrained in the biomedical domain, BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), PubMedBERT (Gu et al., 2020), and BioMedRoBERTa (Gururangan et al., 2020) for the detection of event triggers. The models differ from each other by the corpora in which they were pretrained (all in English), their type of pretraining and the size of their vocabulary, as shown in Table 2.

SciBERT and PubMedBERT were pretrained from scratch using an unique vocabulary and include embeddings that are specific for in-domain words. BioBERT and BioMedRoBERTa were pretrained starting from the BERT checkpoints, their vocabularies are built with general-domain texts (similar to BERT) as well as the initialization of the embeddings.

4.4 Parameters Settings

Experiments are developed using NVIDIA GeForce GTX 1080 Ti (11 GiB) GPU and GeForce RTX 2080 Ti (11 GiB) GPU. The fine-tuning of the transformer models are done with PyTorch, using the Transformers package and the models are taken from Hugging Face [2]. The transformer

[2] https://huggingface.co/

Table 2: Pretrained language models based on transformers used for comparison.

Model	Version	Pretraining	Corpus	Text size	Vocab size
BERT	base uncased	from scratch	WikiPedia + BookCorpus	3.3B words	30,522
BioBERT	base v1.1	from BERT	PubMed	4.5B words	28,996
SciBERT	scivocab cased	from scratch	PMC + Semantic scholar	3.2B words	31,116
PubMedBERT	base uncased	from scratch	PMC + PubMed	3.1B words	30,522
BioMedRoBERTa	base	from BERT	Semantic scholar	7.55B tokens	50,265

models are fine-tuned using the original parameters from BERT, leaving the last layer unfrozen.The training parameters of the classification layers, both linear and Bi-LSTM, are set as follows; batch size of training and test sets of 16, learning rate of 1e-05 and max gradient norm of 10. The maximum length of the sentences is set to 256. All the models are trained during 10 epochs on the training set and evaluated on the test set. The KG is constructed with NetworkX (Hagberg et al., 2008) and the model used to calculate the KG embeddings is implemented with PyTorch, using the base code from (Islam et al., 2021). It is trained using $Adam$ optimizer, with a learning rate of $1e^{-4}$, a weight decay value of 0.01, $pair$ as loss function and 4 as margin value. The training is conducted during 500 epochs, setting 768 as hidden size (embedding dimension) to match with the hidden size of the transformer language model.

5 Results and Discussion

We present a comparison of five different transformer language models for the detection of biomedical trigger events using seven biomedical corpora together. Also, we conduct different strategies for the enrichment of a transformer language model using KG embeddings for the identification of biomedical arguments. We finally compare our results using the test set of the CG dataset with two baselines models from the state-of-the-art.

Table 3 compares the five transformer language models used for the detection of event triggers. For each model is shown the result of adding a linear and a Bi-LSTM classifier. We observe that almost all the models present a better performance when a Bi-LSTM classifier is used, with the exception of BERT, which presents an F1-score of about 2 % higher when using a linear classifier. The highest results are presented in bold, corresponding to SciBERT-*Bi-LSTM*, a model pretrained from scratch using biomedical and general data. PubMedBERT, a model pretrained from scratch using biomedical data, achieves the second best performance when Bi-LSTM is used as classifier, being below SciBERT by around 5 %.

Table 3: Macro-average performance of biomedical event trigger detection (NER) evaluated on the test corpora.

Model	**P**	**R**	**F1**
BERT-*linear*	0.60	0.68	0.64
BERT-*Bi-LSTM*	0.67	0.58	0.62
BioBERT-*linear*	0.52	0.50	0.50
BioBERT-*Bi-LSTM*	0.60	0.56	0.58
SciBERT-*linear*	0.61	0.65	0.63
SciBERT-*Bi-LSTM*	**0.71**	**0.73**	**0.72**
PubMedBERT-*linear*	0.58	0.66	0.61
PubMedBERT-*Bi-LSTM*	0.66	0.69	0.67
BioMedRoBERTa-*linear*	0.52	0.52	0.51
BioMedRoBERTa-*Bi-LSTM*	0.60	0.57	0.58

These last three models, SciBERT, PubMedBERT and BERT, present similar characteristics. They are all pretrained from scratch using very comparable text sizes and have similar vocabulary sizes. However, BERT is pretrained only in the general domain, PubMedBERT in the biomedical domain, and SciBERT in both the general and biomedical domains. The two models that present the lowest performance are BioBERT and BioMedRoBERTa. Both models are pretrained from the BERT weights using biomedical and, biomedical and general data, respectively, and present the largest text sizes of all the models. There is an improvement in both models of around 7 % using a Bi-LSTM classifier compared to a linear classifier. However, their F1-score is around 14 % lower than the one obtained by SciBERT-*Bi-LSTM*.

Table 4 compares our model trained only on the CG corpus with two baseline event extraction models, TEES-CNN (Björne and Salakoski, 2018) and DeepEventMine (Trieu et al., 2020). TEES-CNN is a pipeline model based on a CNN architecture

for event extraction that sequentially applies event trigger detection, argument identification and the construction of events. DeepEventMine is a joint model based on BERT for event extraction that simultaneously detect event triggers, identify arguments and construct the final events. Both models presented state-of-the-art results in the CG task. The results reveal that our proposal achieves better F1-score than TEES-CNN by around 3 % and by DeepEventMine by around 1 %.

Table 4: Comparison of results of biomedical event trigger detection (NER) on the CG corpus.

Model	P	R	F1
TEES-CNN	0.77	0.81	0.79
DeepEventMine	**0.79**	0.83	0.81
SciBERT-*Bi-LSTM* (ours)	0.78	**0.85**	**0.82**

After observing that SciBERT presents the best performance in trigger detection, we enrich its semantic information with the KG of the biomedical corpora for argument identification. We first obtain the KG embeddings of the KG constructed from the seven biomedical corpora. Table 5 compares five KG models used for the computation of these embeddings, TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransD (Ji et al., 2015), DistMult (Yang et al., 2014), SimplE (Kazemi and Poole, 2018). The highest results are presented in bold, obtained with SimplE, which is the model that we use to compute the KG embeddings that are integrated to SciBERT.

Table 5: Results of the KG embeddings using the semantic type information in the triples: *(trigger type, role, argument type)*.

Model	MRR	Hits@1	Hits@3	Hits@10
TransE	0.59	0.26	0.92	0.96
TransH	0.60	0.25	0.96	0.97
TransD	0.60	0.25	0.94	0.96
DistMult	0.86	0.81	0.89	0.96
SimplE	**0.97**	**0.96**	**0.97**	**0.98**

Table 6 compares the results of the different strategies followed to integrate the KG embeddings to SciBERT. SciBERT-$KG_{tr,r,ar}$, which integrates the KG embeddings of the trigger, the role and the argument, presents the highest F1-score. This strategy improves the F1-score by around 18 % compared to when KG embeddings are not integrated. However, we observe that when we add the KG embeddings of the trigger and the argument in SciBERT-$KG_{tr,ar}$, the performance improves by only 1 %, while when we add the KG embeddings of the role in SciBERT-KG_r, the improvement is of 17 %. This reveals that integrating the KG embeddings to SciBERT allows to improve the performance in the identification of biomedical arguments, especially when the KG embeddings of the role are used.

Table 6: Macro-average performance of biomedical argument identification (RE) evaluated on the test corpora.

Model	P	R	F1
SciBERT	0.77	0.70	0.73
SciBERT-$KG_{tr,ar}$	0.80	0.71	0.74
SciBERT-KG_r	**0.93**	0.88	0.90
SciBERT-$KG_{tr,r,ar}$	**0.93**	**0.90**	**0.91**

Table 7 compares our model with TEES-CNN and DeepEventMine for the identification of arguments on the CG corpus. The results reveal that our proposal achieves a better F1 score than TEES-CNN by around 16% and than DeepEventMine by around 15%, which could be significant in reducing event extraction errors related to argument identification.

Table 7: Comparison of results of biomedical argument identification (RE) on the CG corpus.

Model	P	R	F1
TEES-CNN	0.65	0.63	0.64
DeepEventMine	0.63	0.67	0.65
SciBERT-$KG_{tr,r,ar}$ (ours)	**0.87**	**0.75**	**0.80**

6 Conclusions

We analyze different transformer language models for biomedical event trigger detection and argument identification using seven different corpora and KG integration. By comparing the performance of the models, we found that fine-tuning SciBERT with a Bi-LSTM classifier is the best strategy to detect events in different biomedical domains. Furthermore, when SciBERT is enriched with KG embeddings, especially those corresponding to the roles of arguments in events, it significantly improves the identification of biomedical arguments.

Limitations

Our work follows a pipeline approach, where the optimization of event trigger detection and argument identification are done separately. This causes the event trigger detection errors to be passed to the argument identification step. In the future we plan to do the joint optimization of both steps to reduce internally transmitted errors.

Ethics Statement

This work complies with the rules expressed in the ACM Code of Ethics. The NLP application in this work refers to information extraction in the biomedical domain. All the NLP tools and datasets used are mentioned and cited. We do not present a new dataset. The datasets utilized are used as intended. We do not use demographic or identity characteristics information. Our experiments involve around 500 hours of compute time.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Jari Björne and Tapio Salakoski. 2011. Generalizing biomedical event extraction. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 183–191.

Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Yifei Chen. 2021. A transfer learning model with multi-source domains for biomedical event trigger extraction. *BMC genomics*, 22(1):1–18.

Shiyao Cui, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Xuebin Wang, and Jinqiao Shi. 2020. Event detection with relation-aware graph convolutional neural networks. *arXiv e-prints*, pages arXiv–2002.

Soham Dasgupta, Aishwarya Jayagopal, Abel Lim Jun Hong, Ragunathan Mariappan, Vaibhav Rajan, et al. 2021. Adverse drug event prediction using noisy literature-derived knowledge graphs: algorithm development and validation. *JMIR Medical Informatics*, 9(10):e32730.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 352–361.

Arslan Erdengasileng, Qing Han, Tingting Zhao, Shubo Tian, Xin Sui, Keqiao Li, Wanjing Wang, Jian Wang, Ting Hu, Feng Pan, et al. 2022. Pre-trained models, data augmentation, and ensemble learning for biomedical information extraction and document classification. *Database*, 2022.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. *arXiv preprint arXiv:2009.09335*.

Md Kamrul Islam, Sabeur Aridhi, and Malika Smaïl-Tabbone. 2021. Simple negative sampling for link prediction in knowledge graphs. In *International Conference on Complex Networks and Their Applications*, pages 549–562. Springer.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP shared task 2011 workshop*, pages 7–15.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15.

Agustinus Kristiadi, Mohammad Asif Khan, Denis Lukovnikov, Jens Lehmann, and Asja Fischer. 2019. Incorporating literals into knowledge graph embeddings. In *International Semantic Web Conference*, pages 347–363. Springer.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. *arXiv preprint arXiv:2105.13456*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nikola Milošević and Wolfgang Thielemann. 2023. Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, 75:100756.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 1–7.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence*.

David N Nicholson and Casey S Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal*, 18:1414–1428.

Tomoko Ohta, Sampo Pyysalo, and Jun'ichi Tsujii. 2011. Overview of the epigenetics and post-translational modifications (epi) task of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 16–25.

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021. Bert prescriptions to avoid unwanted headaches: A comparison of transformer architectures for adverse drug event detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1740–1747.

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Overview of the infectious diseases (id) task of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 26–35.

Patchigolla VSS Rahul, Sunil Kumar Sahu, and Ashish Anand. 2017. Biomedical event trigger identification using bidirectional recurrent neural network based models. *arXiv preprint arXiv:1705.09516*.

Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5357–5367.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in bert for clinical relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5357–5366.

Javier Sastre, Faisal Zaman, Noirin Duggan, Caitlin McDonagh, and Paul Walsh. 2020. A deep learning knowledge graph approach to drug labelling. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2513–2521. IEEE.

Chen Shen, Hongfei Lin, Xiaochao Fan, Yonghe Chu, Zhihao Yang, Jian Wang, and Shaowu Zhang. 2019. Biomedical event trigger detection with convolutional highway neural network and extreme learning machine. *Applied Soft Computing*, 84:105661.

Laurent Sorber, Marc Van Barel, and Lieven De Lathauwer. 2013. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(l_r,l_r,1) terms, and a new generalization. *SIAM Journal on Optimization*, 23(2):695–720.

Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917.

Meihong Wang, Linling Qiu, and Xiaoli Wang. 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3):485.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. 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 5766–5770.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

SungMin Yang, SoYeop Yoo, and OkRan Jeong. 2020. Denert-kg: Named entity and relation extraction model using dqn, knowledge graph, and bert. *Applied Sciences*, 10(18):6429.

Chrysoula Zerva and Sophia Ananiadou. 2015. Event extraction in pieces: Tackling the partial event identification problem on unseen corpora. In *Proceedings of BioNLP 15*, pages 31–41.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–419.