applied sciences

MDPI

Article

Full-Abstract Biomedical Relation Extraction with Keyword-Attentive Domain Knowledge Infusion

Xian Zhu [1,2], Lele Zhang [3], Jiangnan Du [4] and Zhifeng Xiao [5,*]

1 School of Information Management, Nanjing University, Nanjing 210023, China; zhuxian@njucm.edu.cn
2 School of Health Economics and Management, Nanjing University of Chinese Medicine, Nanjing 210023, China
3 Supply Chain Department, University of Miami, Miami, FL 33136, USA; lxz431@miami.edu
4 Independent Researcher, Shenzhen 518000, China; jiangnan666123@163.com
5 School of Engineering, Penn State Erie, The Behrend College, Erie, PA 16563, USA
* Correspondence: zux2@psu.edu; Tel.: +1-814-898-6252

Abstract: Relation extraction (RE) is an essential task in natural language processing. Given a context, RE aims to classify an entity-mention pair into a set of pre-defined relations. In the biomedical field, building an efficient and accurate RE system is critical for the construction of a domain knowledge base to support upper-level applications. Recent advances have witnessed a focus shift from sentence- to document-level RE problems, which are more challenging due to the need for inter- and intra-sentence semantic reasoning. This type of distant dependency is difficult to understand and capture for a learning algorithm. To address the challenge, prior efforts either attempted to improve the cross-sentence text representation or infuse domain or local knowledge into the model. Both strategies demonstrated efficacy on various datasets. In this paper, a keyword-attentive knowledge infusion strategy is proposed and integrated into BioBERT. A domain keyword collection mechanism is developed to discover the most relation-suggestive word tokens for bio-entities in a given context. By manipulating the attention masks, the model can be guided to focus on the semantic interaction between bio-entities linked by the keywords. We validated the proposed method on the Biocreative V Chemical Disease Relation dataset with an F1 of 75.6%, outperforming the state-of-the-art by 5.6%.

Keywords: relation extraction; chemical-induced disease; pretrained language models; keyword attention; BERT; transformer

Citation: Zhu, X.; Zhang, L.; Du, J.; Xiao, Z. Full-Abstract Biomedical Relation Extraction with Keyword-Attentive Domain Knowledge Infusion. *Appl. Sci.* **2021**, *11*, 7318. https://doi.org/10.3390/app11167318

Academic Editors: Andrzej Sobecki, Higinio Mora, Doina Logofătu and Julian Szymanski

Received: 16 July 2021
Accepted: 6 August 2021
Published: 9 August 2021

1. Introduction

Relation extraction (RE) is a primitive task in natural language processing (NLP). In the context of supervised learning, RE refers to the classification of an entity pair to a set of known relations [1] in a given document or sentence. RE is widely used in biomedical text mining and is usually performed after named entity recognition (NER), jointly discovering and extracting patterns and knowledge from unstructured textual data. Powered by the latest NER and RE algorithms, computers can quickly and accurately identify biomedical entity mentions and the relations between them to build a domain-specific knowledge base to support upper-level applications.

Traditional learning-based methods for RE can be divided into two categories, including feature-based and kernel-based methods [1], which either rely on hand-crafted features or elaborately-designed kernels to perform classification. These methods usually incur error propagation through the learning pipeline, which largely limits the model performance. The rise of deep learning-based models has accelerated the development of a broad spectrum of learning tasks, and RE has also benefited from deep neural models.

One line of efforts takes advantage of the pretrained language models, such as Bidirectional Encoder Representations from Transformers (BERT) [2], Embeddings from Language Models (ELMo) [3], and XLNet [4], which can be fine-tuned for the RE task and present

superior performance. On the other hand, graph neural networks (GNNs) [5–8] have also been extensively investigated in RE due to their intuitive modeling and semantic interpretation capability. In the biomedical field, human annotation is cost-ineffective because of the inaccessible domain knowledge required for annotators. Distant supervised learning [9] was, thus, developed to alleviate the problem and speed up annotation.

Recent RE advances have seen a shift from sentence-level RE to document-level RE. The latter is more challenging due to the inter- and intra-sentence reasoning. In other words, the relation of an entity pair could span multiple sentences, creating a long-range semantic dependency that is hard to detect. Prior efforts attempted to tackle this challenge in two ways: (1) encoding cross-sentence text representations to facilitate distant semantic reasoning [5–7,10] and (2) the infusion of domain or local knowledge into the model to guide training and inference [11–13]. Our study belongs to the latter category. The main hypothesis of this study is that keyword-based domain knowledge can benefit the learning task of document-level RE in the biomedical field.

Our goal is to investigate the role of keywords in RE and their function in performance boosting. To verify this hypothesis, we propose a keyword-attentive knowledge infusion strategy that can be integrated into the BERT neural architecture. The strategy is driven by a custom process of domain keyword collection that aims to discover the most informative tokens that are highly relation-suggestive for bio-entity pairs in a given context. Through the keyword attention masks, the model is guided to focus on the semantic interaction between the bio-entities linked by the keywords. We adopt BioBERT, which has been pretrained on over a million PubMed articles. BioBERT is fine-tuned with the addition of a keyword attention layer for relation classification. Thus, the proposed method is named Kw-BioBERT. Our main contributions are as follows.

- We employ a BERT-based keyword attentive neural architecture, named Kw-BioBERT, for document-level biomedical RE.
- A novel domain keyword collection mechanism is proposed to effectively capture relation-suggestive keywords for knowledge infusion.
- The proposed method is validated on the Biocreative V Chemical Disease Relation (CDR) dataset. The results show that the proposed method outperformed the SOTA by 5.6% in F1 and, thus, can serve as a credible baseline for the CDR dataset.

The rest of this paper is structured as follows. Section 2 covers the prior efforts relevant to this study. Section 3 describes the CDR dataset and the design details of the proposed method. Section 4 provides the implementation details, experimental settings, and results. Section 5 summarizes the work with the limitations and future directions.

2. Related Work

Recent advances in RE have witnessed a wide spectrum of methods and models. In this section, A review of the closely relevant efforts is provided.

2.1. Knowledge Infusion in RE

Knowledge infusion is a common strategy [14,15] to handle low-resource learning tasks with limited supervision. In RE, knowledge infusion has also been found effective. Roy et al. [11] employed the Drug Abuse Ontology (DAO) [16] to determine entity mentions and relations. Similar efforts have appeared in Sousa et al. [12]. In addition to the domain knowledge, local semantic knowledge can also be infused to guide the training. Yu et al. [13] added a position-enhanced module to the BERT neural architecture to encode relative locations between entities. Our proposed method infuses domain knowledge in two ways through (1) biomedical knowledge infused by the pre-trained BioBERT language model and (2) a keyword attentive layer that guides the training to focus on the entity interaction via informative keywords, which, to our best knowledge, has not been seen in prior studies.

2.2. RE Based on Pretrained Language Models

Pretrained Language Models, such as BERT [2], ELMo [3], RoBERTa [17], XLNet [4], T5 [18], and ERNIE [19], have gained explosive attention in numerous NLP tasks, including question and answering [20], named entity recognition [21], summarization [22], text generation [23], and knowledge graph construction [24] due to the strong capability to capture contextual semantic information within text and its self-supervision nature. In the area of RE, Pretrained Language Models have also been extensively utilized, mainly in two ways: (1) to generate contextualized word embeddings [25] and (2) to fine-tune a pretrained model to suit the downstream RE task [26].

We focus on reviewing the second line of work since it is more relevant to our study. Shi et al. [27] proposed a strategy that directly utilizes BERT for RE by only changing the input format to include a document and the entity mentions separately by a [sep] token. Su and Vijay-Shanker [28] proposed a novel fine-tuning process that utilizes all of the outputs from the last transformer layer in BERT, leading to a performance gain. In addition to the base BERT, two of its variants, BioBERT [29], and SciBERT [30] have emerged with a stronger embedding capability to work with scientific publications and gain popularity in the RE task [31–34]. In this work, BioBERT was chosen since it has been pre-trained on over a million PubMed articles, making it very competitive in biomedical NLP tasks.

2.3. Document-Level RE

Document-level RE has recently gained increasing interest in the NLP community since documents provide richer semantic information than sentences. Several datasets that focus on document-level RE have been developed, such as CDR [35,36], DocRED [37], and GDA [38], which have driven the development of innovative models. One line of prior efforts [5–7,10] explored ways to conduct inter- and intra-sentence reasoning [25,39], a major challenge in document-level RE.

Gu et al. [10] employed a maximum entropy (ME) model and a CNN model for inter- and intra-sentence RE, respectively. Bi-affine Relation Attention Network (BRAN) [40] stacks a series of transformers [41] followed by head and tail MLPs and a bi-affine operation that encodes the pairwise token prediction in a 3D tensor. Graph neural networks (GNN) [5–8] have also been a popular choice due to their intuitive modeling ability in RE, where named entities and relations can be modeled as nodes and edges in a graph. Sahu et al. [5] developed a GNN-based model to capture both local and non-local dependency between entity mentions.

Similarly, Wang et al. [6] designed a GNN model, named GLRE, that encodes and aggregates global and local entity and relation representation. Christopoulou et al. [7] proposed an edge-oriented (EoG) GNN that leverages multi-instance learning to enhance intra- and inter-sentence reasoning. Compared to the prior studies that focused on modeling and reasoning, our work focuses on domain knowledge infusion, which has not been extensively explored in the field of document-level RE. One relevant work is by Sousa et al. [12], which injected domain ontology knowledge into the model, resulting in performance gains. On the other hand, our work investigates the role of keywords, which is a novel method of knowledge infusion.

3. Materials and Methods

3.1. Datasets

The CDR dataset [35,36] was adopted in this work to evaluate the proposed model. The CDR dataset models the chemical–disease relations, namely chemical-induced disease (CID) relations, and is created at the abstract level with entity-linked mention annotations, which are featured by long-range and cross-sentence relations. Specifically, A CID relation marked in the dataset could be either a putative mechanistic relation or a biomarker relation. The former means that the chemical is a potential etiology of the disease (e.g., cancer x is caused by exposure to chemical y); the latter, on the other hand, indicates a correlation

between the chemical and the disease (e.g., an increased abundance of chemical X in the brain correlates with disease Y).

The two relation sub-types are treated as a unified CID relation, creating a binary classification problem, i.e., CID/non-CID relation. According to [35], the development of the dataset involved four annotators with medical training background. A double-annotation strategy was adopted; namely, each abstract was independently labeled by two annotators. The dispute was resolved by a third and senior annotator. All annotations were performed using PubTator [42].

Table 1 shows the statistical information of the CDR dataset, where 1500 abstracts are equally divided into training, development, and test sets. The mentions of diseases and chemicals are about equally distributed into the three sets of data. The size of positive samples, namely the chemical–disease (CD) pairs that present a CID relation, is 3116, which is about one-fourth of the size of negative samples, i.e., the CD pairs without a CID relation. The imbalanced distribution of classes brings difficulty to both training and evaluation. In addition to the original training set shown in Table 1, the task contains an additional training set [43] of 15,448 weakly labeled PubMed abstracts with 26,657 positive CID relations and 146,057 negative ones. This extra data is used as a secondary source for training.

Table 1. Stats for the CDR dataset.

Data Split	Abstracts	Diseases	Chemicals	Pos.	Neg.
Training	500	4182	5203	1038	4280
Dev.	500	4244	5347	1012	4136
Test	500	4424	5385	1066	4270
Total	1500	12,850	15,935	3116	12,686

Table 2 displays an abstract sample with annotations in the CDR dataset. The first two sections are the original article title and abstract. The third section lists the entity mentions, where each row follows a format of "PMID offset length mention_text entity entity_ID", which describes an entity mention with an exact location. The last section lists the CID relations that follow a format of "PMID relation_type head_entity tail_entity".

Table 2. A sample with annotation in the CDR dataset.

Title	20633755 | t | Suxamethonium induced prolonged apnea in a patient receiving electroconvulsive therapy.
Abstract	20633755 | a | Suxamethonium caused prolonged apnea in patients in whom pseudocholinesterase enzyme gets deactivated by organophosphorus (OP) poisons. Here, we present a similar incident in a severely depressed patient who received electroconvulsive therapy (ECT). Prolonged apnea, in our case, ensued because the information about a suicidal attempt by OP compound was concealed from the treating team.
Entity Mentions	20633755 0 13 Suxamethonium Chemical D013390 20633755 32 37 apnea Disease D001049 20633755 88 101 Suxamethonium Chemical D013390 20633755 119 124 apnea Disease D001049 20633755 193 222 organophosphorus (OP) poisons Chemical D009943 20633755 274 283 depressed Disease D003866 20633755 348 353 apnea Disease D001049 20633755 423 434 OP compound Chemical D009943
Relations	20633755 CID D009943 D001049 20633755 CID D013390 D001049

3.2. Learning Problem

Given the CDR dataset, the learning task can be formulated as follows. Let D_{train}, D_{dev}, and D_{test} denote the training, development, and test set, respectively. Each instance in the dataset can be defined as (x_i, y_i), where $x_i = (A_k, C_m, D_n)$, representing a CD pair (C_m, D_n) that appears in an abstract A_k (i, k, m, and n are indices), and y_i is a binary target; where 1 indicates a CID relation of the CD pair, i.e., a positive instance, and 0 otherwise. The learning problem is to develop a model that takes $x_i \in D_{test}$ as input and makes a prediction $\hat{y}_i$ that should approximate the ground truth y_i as much as possible. It is noted that the problem belongs to document-level RE, where the head and tail entity mentions could span across multiple sentences in the abstract.

3.3. System Framework

The system framework is given in Figure 1. First, the BioBERT base model is fine-tuned using the training set. The tuned BioBERT model is used for keyword extraction, generating a collection of seed keywords that are highly relation-suggestive. The seed keyword set is then expanded to form the final domain-specific set of keywords. We modify the BioBERT network by adding a keyword-attentive layer in parallel with the last transformer layer, similar to [44]. The resulting Kw-BioBERT model is then fine-tuned on the training set, with the keywords injected into the model as external domain knowledge. The tuned Kw-BioBERT is evaluated on the test set to obtain the final result.

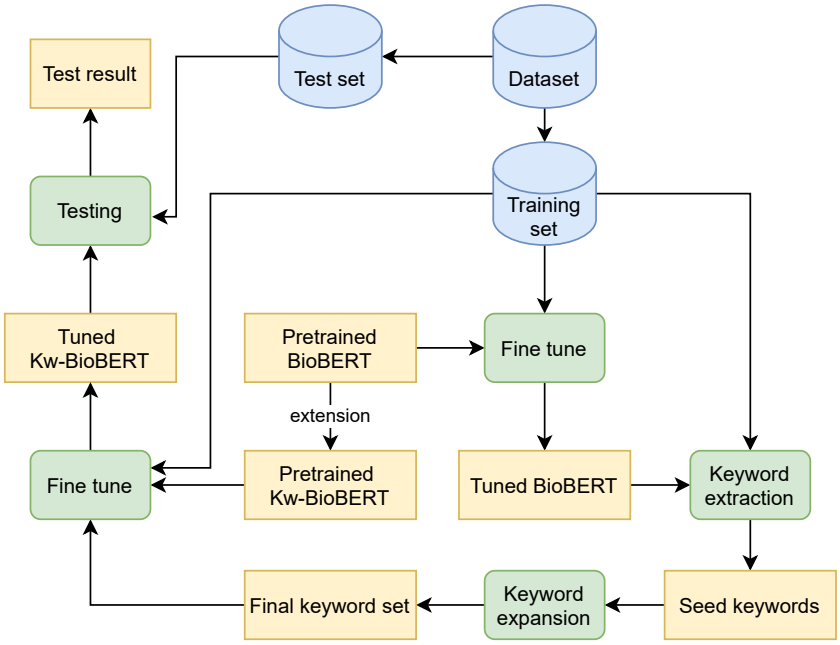

Figure 1. System framework. BioBERT is fine-tuned on the CDR training set. Then, the keyword extraction algorithm is applied to the tuned BioBERT model to generate a set of seed keywords, expanded to form the final keyword set. The BioBERT is changed to Kw-BioBERT and further tuned on the training set with the keyword attention mechanism enabled. Finally, the tuned Kw-BioBERT is evaluated on the CDR test set.

3.4. Network Architecture

A network architecture (as shown in Figure 2) similar to [44] is adopted. However, our version has two differences compared to the original design, including the input form and keyword manipulation. The input is a sequence pair (seq_A, seq_B), where seq_A starts with a [cls] token, ends with a [sep] token, and has a tokenized full abstract A in the middle; seq_B specifies the head and tail entity mentions. For the CDR dataset, seq_B consists of a chemical entity mention C (the head) and a disease entity mention D (the tail), both appearing in A. The model's job is to understand the semantic relation between C and D,

given A as a context. The neural architecture is modified from BioBERT (with the same neural architecture with BERT), adding a keyword attentive layer side by side with the last transformer encoder layer in BERT.

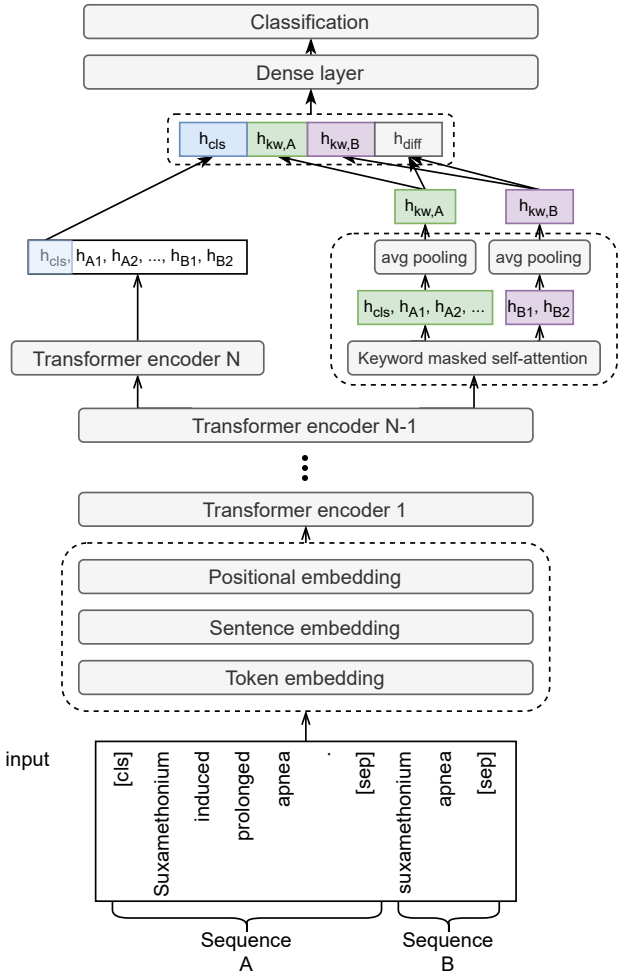

Figure 2. Neural architecture of Kw-BioBERT. A keyword-attentive layer is added in parallel with the last transformer encoder to represent the semantic interaction between the relation-suggestive tokens and entity mentions.

The keyword attentive layer differs from a transformer encoder in two aspects. First, the attention masks, in the transformer encoder, are used to mask the padding tokens so that they do not participate attention; in the keyword attentive layer, however, the attention masks are also manipulated to allow tokens in seq_A to only attend the two entity mentions in seq_B and allow the two tokens in seq_B to only attend the keywords in seq_A. In other words, for each token in seq_A, we only care about its attention (or impact) on the two entity mentions in seq_B; also, for each token in seq_B, we only consider its attention on the keywords in seq_A.

With the manipulation attention masks, the model learns how the entity mentions and the keywords interact and jointly determine the relation. The output of the keyword attentive layer is a vector of hidden states $[h_{cls}, h_{A1}, h_{A2}, ..., h_{B1}, h_{B2}, h_{sep}]$, which has the same size as the input. The hidden state vector can be divided to two sections corresponding to seq_A and seq_B. Then, a pooling operation is applied to each section individually, producing $h_{kw,A}$ and $h_{kw,B}$, which represent the aggregated and keyword-attentive embeddings for seq_A and seq_B, respectively. Now, the semantic difference between $h_{kw,A}$ and $h_{kw,B}$ is denoted as h_{diff}, which is defined as

$$h_{diff} = [h_{kw,A} - h_{kw,B}; h_{kw,B} - h_{kw,A}] \quad (1)$$

where $[;]$ is the concatenation operation. Next, the four pieces of information are concatenated, including the h_{cls} from the last transformer encoder, $h_{kw,A}$, $h_{kw,B}$, and h_{diff}, and feed the resulting vector into the detection head, which consists of a dense layer and a softmax function.

3.5. Keywords Collection

The mission keywords collection is to discover the word tokens that are relation-suggestive. For instance, in "Suxamethonium induced prolonged apnea in a patient receiving electroconvulsive therapy.", the two entity mentions "Suxamethonium" and "apnea" are linked via the verb "induced", making it a keyword that suggests a CID relation. In another example, "Myasthenia gravis presenting as weakness after magnesium administration.", the dominate keyword is not apparent, and words "presenting", "after", and "administration" jointly affect the CID relation between "magnesium" and "Myasthenia gravis".

The process of identifying the seed keywords is as follows. For each positive CD pair in each abstract within the CDR training set, we do the following: the instance (A, C, D) is fed into BERT that has been tuned on the training set and obtains a prediction. If the prediction does not match the ground truth, we move on to the next CD pair; otherwise, the abstract is scanned token by token; specifically, for each token that is not a (1) entity mention, (2) punctuation, or (3) stop word, we mask it in the abstract and obtain A_{masked}; then, (A_{masked}, C, D) is fed into BERT again and we record a change in the output confidence.

The top three tokens that cause the most confidence drop are kept and added into the candidate keyword set. The rationale is that if a token, masked in the abstract, leads to a significant confidence drop, it means that the token is highly relation-suggestive for the CD pair, given the abstract as a context. This way, a set of candidate keywords is collected and further manually selected to form a seed keyword set. Examples of these keywords include "induced", "statistically", "maintenance", "consumption", and "idiopathic", etc. More keyword examples are provided in Section 4.2.

To enhance the keyword diversity, the synonyms of the seed keywords are added into the keyword set. For instance, the word "induce" is semantically close to "produce", "cause", "effect", and "provoke", and could be used interchangeably when describing a CID relation. After adding the synonyms, the final keyword set is created and ready for use.

3.6. Keyword-Attentive Knowledge Infusion

We take advantage of the attention mask feature implemented in BERT. Figure 3a shows a positive instance with a chemical entity mention (CEM) "Suxamethonium" and a disease entity mention (DEM) "apnea", linked by a keyword "induced". To ensure that each token in seq_A only participates attention to the two tokens in seq_B and that each token in seq_B only participates attention to the keyword tokens in seq_A, we employ a binary matrix, denoted by M_{AttMsk}, of size $l \times l$. Let l_A and l_B denote the length of seq_A and seq_B, respectively.

We then have $l = l_A + l_B$. Generally, the ith row in M_{AttMsk} specifies how token i of the input attends other tokens. In particular, $M_{AttMsk}(i, j) = 1$ indicates that token i attends token j, giving a one at row i and column j in the matrix; also, $M_{AttMsk}(i, j) = 0$ means that token i does not attend token j, posting a zero at position (i, j) of the matrix. To fulfill our needs, for all tokens in seq_A, a common attention mask vector, with all zeros in the first l_A positions and two ones in the two positions corresponding to the entity mentions in seq_B, can serve the purpose. On the other hand, all tokens in seq_B share an attention mast vector, with all zeros in all positions except the ones where the keywords reside. Figure 3b shows a complete example of M_{AttMsk}, given the input in Figure 3a.

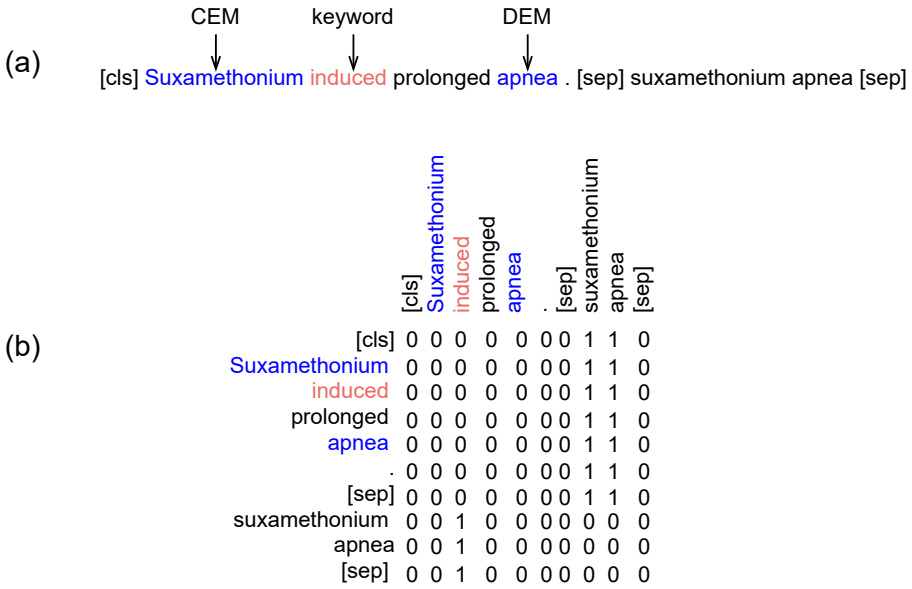

Figure 3. Self-attention masks. (**a**) A sample with a CID relation. (**b**) The attention mask matrix associated with the sample in (**a**). Each token i is assigned with a binary vector (i.e., row i in M_{AttMsk}) that specifies which tokens it attends to. If the jth element at row i is zero, token j is not attended by i; otherwise, j is attended by i.

4. Evaluation

4.1. Training Setting

The proposed model and keyword collection procedure were implemented using Python 3.6.10 and TensorFlow 1.13. Experiments were conducted on an Nvidia V100. On the CDR training set, each epoch took about 14 min (for Kw-BioBERT); on the additional training data, the running time per epoch was about 182 min. Two hyperparameters were tuned, including the number of transformer layers and the training epochs, with a learning rate of 3×10^{-5}. The results are reported in the following sections.

4.2. Keywords

As described in Section 3.5, for each positive instance, when masked and fed into the BERT model, the top three tokens that caused the largest confidence drop were recorded and considered as candidate keywords. The process was applied to the CDR training set, and three token sets that store the top-three relation-suggestive tokens were obtained. Figure 4 selectively displays the seed keywords discovered in the process, sorted by frequency in a decreasing order. Subfigures (a), (b), and (c) correspond to the tokens resulting in the most, second-most, and third-most confidence drop. In our experiment, a total of 1736 candidate keywords is identified. After a round of manual selection, 235 tokens remain to form the seed keyword set, which is further expanded to a keyword set of 943 tokens, with the synonyms (found through WordNet [45]) added.

4.3. Performance Metric

Due to the imbalanced class distribution, accuracy is not adequate, because it may drive the learning algorithm to classify all instances to the major class. For our case, this could yield numerous false negatives, meaning that the positive CID relations are not detected. Thus, F1 is adopted as the main performance metric for model evaluation since F1 is superior to accuracy in the case of imbalanced class distribution. We also report precision (Pre), which reflects the number of false alarms, and recall (Rec), which implies the number of missed CID relations. Intuitively, the higher the precision, the fewer the false alarms; also, the the higher the recall, the fewer the missed CID relations. In addition, the Pre-Rec gap should be monitored: if the gap is too large, it means that the model focuses

on the optimization of a single metric, rather than both, which should be avoided. With the given true positives (TP), true negatives (TN), and false positives (FP), the definitions of Pre, Rec, and F1 can be given below.

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{2}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \times 100\% \tag{4}$$

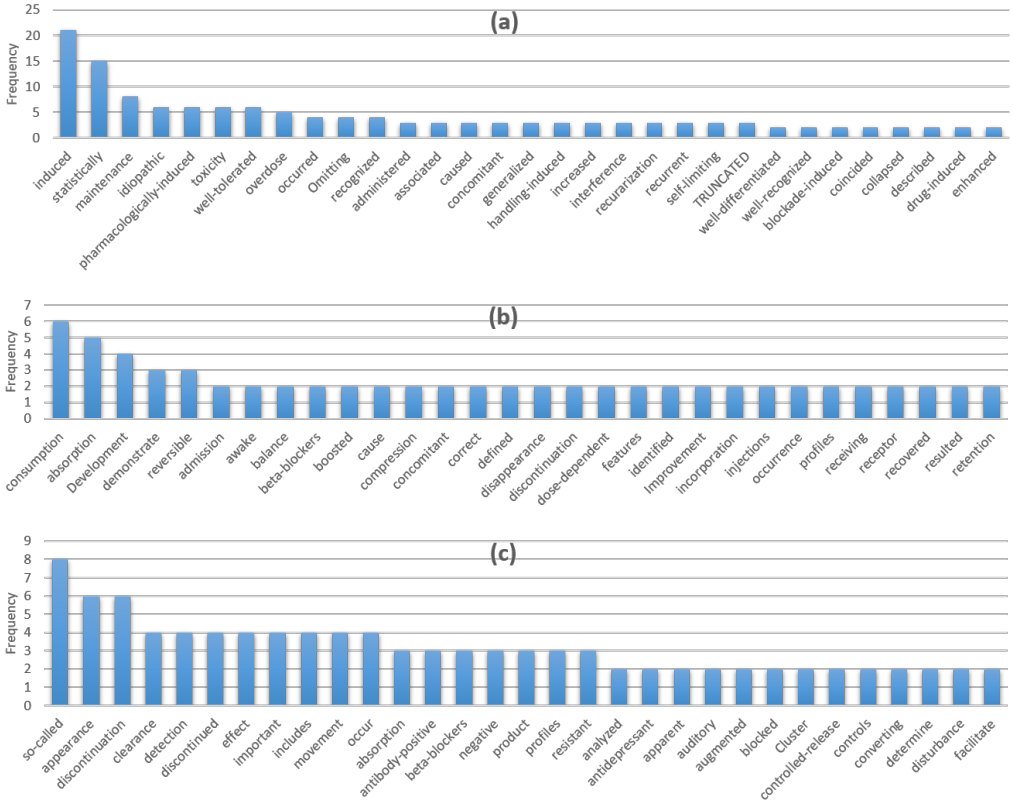

Figure 4. Keyword frequency. Subfigures (**a**–**c**) correspond to the tokens resulting in the most, second-most, and third-most confidence drop.

4.4. Benchmark

We selected the following models that were evaluated on the CDR test set. The performance results of these models are quoted from the original papers.

- Gu et al. [10] adopted a maximum entropy (ME) model and a CNN model for inter- and intra-sentence RE, respectively.
- Bi-affine Relation Attention Network (BRAN) [40] consists of a stack of modified transformers, a head and tail MLP, and a bi-affine operation to output a 3D tensor that models pairwise token relations.
- Sahu et al. [5] employed a GCNN to model entities and their relations in a document and demonstrated that GCNN can capture both local and non-local dependency, which helps to boost performance.
- Christopoulou et al. [7] proposed an edge-oriented graph (EoG) neural model to learn intra- and inter-sentence via multi-instance learning.
- Nan et al. [39] developed a latent structure refinement strategy that allows reasoning across sentences and automated latent graph construction.

- Sousa et al. [12] proposed to utilize external domain-specific ontologies to enhance the performance of biomedical RE. The proposed system, named BiOnt, injects additional knowledge into the model, leading to performance gain.
- Wang et al. [6] developed a graph-based neural model named GLRE that encodes and aggregates global and local entity and relation representation for document-level RE.
- Zeng et al. [25] designed a neural architecture named SIRE that can represent intra- and inter-sentential relations. In addition, SIRE is featured with a novel logical reasoning module that covers more reasoning chains compared to the prior efforts. SIRE posts the highest F1 on the CDR dataset among all of the investigated studies; thus, SIRE represents the SOTA.

4.5. Key Design Choices

Two hyperparameters are tuned.

- For the number of transformer layers, 2, 4, 6, 8, 10, and 12 layers were tested. Each experiment ran for five epochs. As shown in Table 3, the performance of Kw-BioBERT with twelve transformer layers was the best, with an F1 of 75.8%.
- For the training epochs, we reported training and test performance with 1, 2, through 5 epochs. Since BioBERT has been pre-trained, the effort of fine-tuning a medium sized dataset can be greatly reduced. In our experiments, the test F1 started to stabilize after the first epoch and reached a peak at the third epoch, with an F1 of 76.4%, as shown in Table 4. It is also noted that performance gap between training and test F1, indicating overfitting, which can usually be addressed by an increase of training data.

Table 3. Number of transformer layers vs. performance. The highest value of each metric is marked in bold.

# transformers	Pre	Rec	F1
Kw-BioBERT-2	55.3	58.2	56.7
Kw-BioBERT-4	68.3	69.3	68.8
Kw-BioBERT-6	73.4	75.7	74.5
Kw-BioBERT-8	74.5	75.2	74.9
Kw-BioBERT-10	74.3	75.2	74.7
Kw-BioBERT-12	**74.6**	**77.0**	**75.8**

Table 4. Training epochs vs. performance. The highest value of each metric is marked in bold.

	# epochs	1	2	3	4	5
Training	Pre	87.6	94.4	97.4	98.5	**98.8**
	Rec	89.1	94.2	97.2	98.5	**99.0**
	F1	88.3	94.3	97.3	98.5	**98.9**
Test	Pre	72.2	**75.6**	75.5	75.1	74.6
	Rec	75.6	76.6	**77.5**	77.5	77.0
	F1	73.6	76.1	**76.4**	76.2	75.8

4.6. Ablation Study

Table 5 shows the result of an ablation study, in which four models are evaluated, including BERT, Kw-BERT, BioBERT, and Kw-BioBERT. There are two primary observations. First, adding a keyword attention layer to the base models brought a performance gain of about two points, with a 2.1-point gain (in F1) on BERT and a 1.9-point gain on BioBERT. Second, switching BERT to BioBERT brought a gain of around eight points, by looking at BERT vs. BioBERT, and Kw-BERT vs. Kw-BioBERT. This gain is surprising but explainable since BioBERT is pretrained on corpora in the biomedical domain at a large scale; thus, BioBERT can better encode and represent the semantic meaning of PubMed abstracts.

This experiment also validates the efficacy of the proposed method of keyword-attentive knowledge infusion, which nicely complements the pretrained language models. Essentially, the utilization of BioBERT and keywords can be both regarded as knowledge infusion, but at two levels. BioBERT receives domain knowledge by pretraining, which is self-supervised; the keyword-attentive layer, on the other hand, injects task-specific knowledge (i.e., relation-suggestive tokens and semantic interaction between bio-entities) during training, which is supervised.

Table 5. Ablation study. The highest value of each metric is marked in bold.

	Pre	Rec	F1
BERT	65.4	67.9	66.3
Kw-BERT	68	68.9	68.4
BioBERT	73.6	75.6	74.5
Kw-BioBERT	**75.5**	**77.5**	**76.4**

4.7. Comparison with the Benchmarks

We present the performance of the benchmarks and the proposed Kw-BioBERT in Table 6 on the CDR test set. We observed that Kw-BioBERT outperformed the SOTA, namely SIRE, by 5.6% in F1. When trained on additional data (denoted by model + data in the last two rows of the table), our method posted an F1 of 80.8%, outperforming Bran by 14.6%. The latter has been used as a credible baseline in many prior studies. In addition, the Pre-Rec gap of our model is only two points, which is smaller than that of other benchmarks listed in the table, e.g., GLRE (7.1%), EoG(3.1%), and Bran (15.2%), further demonstrating the superiority of our model that seeks for optimizing both Pre and Rec.

Table 6. The performance of the benchmarks and our method on the CDR test set. The highest value of each metric is marked in bold.

Model	Year	Pre	Rec	F1
ME+CNN [10]	2017	55.7	68.1	61.3
Bran [40]	2018	55.6	70.8	62.1
GCNN [5]	2019	52.8	66	58.6
EoG [7]	2019	62.1	65.2	63.6
LSR w/o MDP Nodes [39]	2020	-	-	64.8
BiOnt [12]	2020	57.7	71.7	64
GLRE [6]	2020	65.1	72.2	68.5
SIRE [25] (SOTA)	2021	-	-	70.8
Kw-BioBERT (ours)	2021	**75.5**	**77.5**	**76.4**
Bran + data	2018	64.0	69.2	66.2
Kw-BioBERT + data	2021	**82.9**	**79.2**	**80.8**

4.8. Overhead of the Keyword Attention Mechanism

Table 7 reports an overhead comparison between Kw-BioBERT and BioBERT in terms of the training and inference speed, both in examples per second (ex/s). BioBERT posted an average speed of 11 ex/s during training, and Kw-BioBERT was almost two times faster, with a speed of 20.5 ex/s. During inference, the speeds of BioBERT and Kw-BioBERT were 66.5 and 72.6 ex/s, respectively. The increase of speed brought by Kw-BioBERT is mainly due to the attention layer added into BERT and replacing the Nth standard transformer encoder. In other words, the original design of self-attention performs pair-wise attention between every pair of tokens, and the proposed keyword masked self-attention only concerns the attention (1) from tokens in seq_A to the two entity mentions in seq_B and (2) from the two mentions in seq_B to the keywords in seq_A, greatly reducing the attention calculations.

Table 7. Overhead comparison between Kw-BioBERT and BioBERT.

	Training Speed (ex/s)	Inference Speed (ex/s)
BioBERT	11	66.5
Kw-BioBERT	20.5	72.6

5. Conclusions

Document-level RE has been given increased research attention recently. A broad spectrum of deep models has been explored, including CNN, GNN, and transformer-based models. To address the challenge of distant dependency reasoning, there are two lines of efforts. The first category focuses on improving cross-sentence representation, and GNNs become an intuitive modeling choice due to their straightforward way of representing entities and relations as nodes and edges, facilitating long-range reasoning. The second line of studies, on the other hand, explores the utilization of external knowledge. Our work in this study belongs to the second category.

To verify the hypothesis that domain keywords can improve the model performance for the document-level RE task, a keyword-attentive knowledge infusion strategy was proposed. We developed a custom process of domain keyword collection to identify and store the highly relation-suggestive tokens in a given document. By manipulating the attention masks, these keywords were injected into BERT to guide the learning algorithm to focus on the semantic interaction between the bio-entities linked by the keywords.

In addition, we adopt BioBERT, a BERT variant pretrained on over a million PubMed articles, for fine-tuning. These joint efforts brought together created a model with superior performance, outperforming the SOTA by 5.6%, on the CDR dataset. Thus, we concluded that the hypothesis of this study can be accepted, and the goal was achieved. The new-high F1 score indicates that the proposed Kw-BioBERT can serve as a credible benchmark of the CDR dataset for future research.

This study has the following limitations, which will be addressed in future work. First, the imbalanced sample distribution issue brought difficulties in training an accurate model, which is a common issue for most RE datasets. We plan to adopt positive instance sampling or augmentation techniques to rebalance the samples in different classes. Second, it will be of interest to evaluate the proposed Kw-BioBERT to other document-level RE datasets, with more types of bio-entities and relations. In an ongoing study, a novel biomedical RE dataset is being developed, with five types of entities and six types of relations. The proposed Kw-BioBERT will be evaluated on this new dataset. Lastly, we studied neither the impact of keywords quality on the model performance nor the alternatives of domain knowledge infusion, which are worthy of further investigation.

Author Contributions: Conceptualization and methodology, X.Z., L.Z., J.D. and Z.X.; software, validation, and original draft preparation, X.Z., L.Z. and J.D.; review and editing, X.Z. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting the conclusions of this article is available at https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/ (accessed on 2 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, S. A survey of deep learning methods for relation extraction. *arXiv* **2017**, arXiv:1705.03645.
2. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

3. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
4. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
5. Sahu, S.K.; Christopoulou, F.; Miwa, M.; Ananiadou, S. Inter-sentence relation extraction with document-level graph convolutional neural network. *arXiv* **2019**, arXiv:1906.04684.
6. Wang, D.; Hu, W.; Cao, E.; Sun, W. Global-to-local neural networks for document-level relation extraction. *arXiv* **2020**, arXiv:2009.10359.
7. Christopoulou, F.; Miwa, M.; Ananiadou, S. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. *arXiv* **2019**, arXiv:1909.00228.
8. Wang, J.; Chen, X.; Zhang, Y.; Zhang, Y.; Wen, J.; Lin, H.; Yang, Z.; Wang, X. Document-level biomedical relation extraction using graph convolutional network and multihead attention: Algorithm development and validation. *JMIR Med. Inform.* **2020**, *8*, e17638. [CrossRef]
9. Feng, X.; Guo, J.; Qin, B.; Liu, T.; Liu, Y. Effective Deep Memory Networks for Distant Supervised Relation Extraction. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; Volume 17.
10. Gu, J.; Sun, F.; Qian, L.; Zhou, G. Chemical-induced disease relation extraction via convolutional neural network. *Database* **2017**, *2017*, bax024. [CrossRef]
11. Roy, K.; Lokala, U.; Khandelwal, V.; Sheth, A. "Is depression related to cannabis?": A knowledge-infused model for Entity and Relation Extraction with Limited Supervision. *arXiv* **2021**, arXiv:2102.01222.
12. Sousa, D.; Couto, F.M. BiOnt: Deep learning using multiple biomedical ontologies for relation extraction. *Adv. Inf. Retr.* **2020**, *12036*, 367.
13. Yu, H.; Cao, Y.; Cheng, G.; Xie, P.; Yang, Y.; Yu, P. Relation extraction with BERT-based pre-trained model. In Proceedings of the IEEE 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15–19 June 2020; pp. 1382–1387.
14. Peng, H.; Ning, Q.; Roth, D. Knowsemlm: A knowledge infused semantic language model. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 550–562.
15. He, Y.; Zhu, Z.; Zhang, Y.; Chen, Q.; Caverlee, J. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. *arXiv* **2020**, arXiv:2010.03746.
16. Cameron, D.; Smith, G.A.; Daniulaityte, R.; Sheth, A.P.; Dave, D.; Chen, L.; Anand, G.; Carlson, R.; Watkins, K.Z.; Falck, R. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *J. Biomed. Inform.* **2013**, *46*, 985–997. [CrossRef] [PubMed]
17. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
18. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683.
19. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129.
20. Wang, Z.; Ng, P.; Ma, X.; Nallapati, R.; Xiang, B. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv* **2019**, arXiv:1908.08167.
21. Liang, C.; Yu, Y.; Jiang, H.; Er, S.; Wang, R.; Zhao, T.; Zhang, C. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 23–27 August 2020; pp. 1054–1064.
22. Liu, Y. Fine-tune BERT for extractive summarization. *arXiv* **2019**, arXiv:1903.10318.
23. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
24. Liu, W.; Zhou, P.; Zhao, Z.; Wang, Z.; Ju, Q.; Deng, H.; Wang, P. K-bert: Enabling language representation with knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 2901–2908.
25. Zeng, S.; Wu, Y.; Chang, B. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction. *arXiv* **2021**, arXiv:2106.01709.
26. Lin, C.; Miller, T.; Dligach, D.; Bethard, S.; Savova, G. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 65–71.
27. Shi, P.; Lin, J. Simple bert models for relation extraction and semantic role labeling. *arXiv* **2019**, arXiv:1904.05255.
28. Su, P.; Vijay-Shanker, K. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea, 16–19 December 2020; pp. 2522–2529.
29. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]

30. Beltagy, I.; Lo, K.; Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv* **2019**, arXiv:1903.10676.
31. Alimova, I.; Tutubalina, E. Multiple features for clinical relation extraction: A machine learning approach. *J. Biomed. Inform.* **2020**, *103*, 103382. [CrossRef] [PubMed]
32. Li, D.; Ji, H. Syntax-aware multi-task graph convolutional networks for biomedical relation extraction. In Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), Hong Kong, China, 3 November 2019; pp. 28–33.
33. Giles, O.; Karlsson, A.; Masiala, S.; White, S.; Cesareni, G.; Perfetto, L.; Mullen, J.; Hughes, M.; Harland, L.; Malone, J. Optimising biomedical relationship extraction with biobert: Best practices for data creation. *bioRxiv* **2020**. [CrossRef]
34. Wei, Q.; Ji, Z.; Si, Y.; Du, J.; Wang, J.; Tiryaki, F.; Wu, S.; Tao, C.; Roberts, K.; Xu, H. Relation extraction from clinical narratives using pre-trained language models. In Proceedings of the AMIA Annual Symposium Proceedings, Washington, DC, USA, 16–20 November 2019; American Medical Informatics Association: Bethesda, MD, USA, 2019; Volume 2019, p. 1236.
35. Li, J.; Sun, Y.; Johnson, R.J.; Sciaky, D.; Wei, C.H.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Wiegers, T.C.; Lu, Z. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database* **2016**, *2016*, baw068. [CrossRef]
36. Wei, C.H.; Peng, Y.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Li, J.; Wiegers, T.C.; Lu, Z. Assessing the state of the art in biomedical relation extraction: Overview of the BioCreative V chemical–disease relation (CDR) task. *Database* **2016**, *2016*, baw032. [CrossRef] [PubMed]
37. Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; Sun, M. DocRED: A large-scale document-level relation extraction dataset. *arXiv* **2019**, arXiv:1906.06127.
38. Wu, Y.; Luo, R.; Leung, H.C.; Ting, H.F.; Lam, T.W. Renet: A deep learning approach for extracting gene-disease associations from literature. In Proceedings of the International Conference on Research in Computational Molecular Biology, Washington, DC, USA, 5–8 May 2019; Springer: Cham, Switzerland, 2019; pp. 272–284.
39. Nan, G.; Guo, Z.; Sekulić, I.; Lu, W. Reasoning with latent structure refinement for document-level relation extraction. *arXiv* **2020**, arXiv:2005.06312.
40. Verga, P.; Strubell, E.; McCallum, A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *arXiv* **2018**, arXiv:1802.10569.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
42. Wei, C.H.; Kao, H.Y.; Lu, Z. PubTator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41*, W518–W522. [CrossRef] [PubMed]
43. Peng, Y.; Wei, C.H.; Lu, Z. Improving chemical disease relation extraction with rich features and weakly labeled data. *J. Cheminform.* **2016**, *8*, 53. [CrossRef] [PubMed]
44. Miao, C.; Cao, Z.; Tam, Y.C. Keyword-Attentive Deep Semantic Matching. *arXiv* **2020**, arXiv:2003.11516.
45. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]