symmetry

MDPI

Article

EMSI-BERT: Asymmetrical Entity-Mask Strategy and Symbol-Insert Structure for Drug–Drug Interaction Extraction Based on BERT

Zhong Huang [1,2,3], Ning An [1,3,*], Juan Liu [2,*] and Fuji Ren [4]

[1] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, School of Computer Science and Information, Hefei University of Technology, Hefei 230009, China
[2] School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University, Anqing 246011, China
[3] Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230601, China
[4] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610056, China
* Correspondence: ning.g.an@acm.org (N.A.); juanl@aqnu.edu.cn (J.L.)

Abstract: Drug-drug interaction (DDI) extraction has seen growing usage of deep models, but their effectiveness has been restrained by limited domain-labeled data, a weak representation of co-occurring entities, and poor adaptation of downstream tasks. This paper proposes a novel EMSI-BERT method for drug–drug interaction extraction based on an asymmetrical Entity-Mask strategy and a Symbol-Insert structure. Firstly, the EMSI-BERT method utilizes the asymmetrical Entity-Mask strategy to address the weak representation of co-occurring entity information using the drug entity dictionary in the pre-training BERT task. Secondly, the EMSI-BERT method incorporates four symbols to distinguish different entity combinations of the same input sequence and utilizes the Symbol-Insert structure to address the week adaptation of downstream tasks in the fine-tuning stage of DDI classification. The experimental results showed that EMSI-BERT for DDI extraction achieved a 0.82 F1-score on DDI-Extraction 2013, and it improved the performances of the multi-classification task of DDI extraction and the two-classification task of DDI detection. Compared with baseline Basic-BERT, the proposed pre-training BERT with the asymmetrical Entity-Mask strategy could obtain better effects in downstream tasks and effectively limit "Other" samples' effects. The model visualization results illustrated that EMSI-BERT could extract semantic information at different levels and granularities in a continuous space.

Keywords: drug–drug interaction; BERT; entity-mask strategy; symbol-insert structure; symmetry and asymmetry; machine learning

Citation: Huang, Z.; An, N.; Liu, J.; Ren, F. EMSI-BERT: Asymmetrical Entity-Mask Strategy and Symbol-Insert Structure for Drug–Drug Interaction Extraction Based on BERT. *Symmetry* **2023**, *15*, 398. https://doi.org/10.3390/sym15020398

Academic Editor: José Carlos R. Alcantud

Received: 15 December 2022
Revised: 31 January 2023
Accepted: 31 January 2023
Published: 2 February 2023

1. Introduction

Drug–drug interaction (DDI) refers to extracting the relation of the combination and interaction of two or more drugs in the human body, which is the most common task in the field of biomedical relation extraction [1,2]. Recent studies have shown that a growing number of people need to take multiple drugs simultaneously, and the interaction among these drugs could severely affect their health [3]. Further understanding of drug–drug interaction and designing a DDI classification system is critical to reducing drug abuse or accidents [4]. As a result, researchers have paid increasing attention to DDI-related work [3,5,6]. Various DDI databases, including DrugBank [7], Mint [8], and IntAct [9], have emerged. Meanwhile, the literature on biological drug mechanisms has dramatically increased as biomedicine develops rapidly [5]. DDI extraction is likely to no more be analyzed and discovered manually. Therefore, an automatic and accurate DDI extraction system using the massive unlabeled literature data on drug mechanisms is highly desired [10,11].

Simultaneously taking multiple drugs has caused frequent human body injuries and drug abuse accidents [2,12,13]. Exploring drug–drug interactions through various medical diaries or records has gained significant research interest recently. Early DDI extraction methods are mainly based on traditional machine learning techniques typified by feature engineering [1,2]. These methods are poorly suited to relation extraction due to the limitations of sparse data and templates. In addition, they are also challenging to extend due to the non-reusability of the feature extraction strategies. Recently, the methods based on deep learning have been rapidly applied to DDI extraction, especially pre-training models typified by Bidirectional Encoder Representations from Transformers (BERT), garnering widespread attention in the field of relation extraction [14–16]. However, how to improve the pre-training effect of BERT and better adapt to downstream DDI classification tasks still needs to be further explored. Furthermore, although massive amounts of data describing drug mechanisms are available at present, the DDI annotation requires a large amount of medical professional knowledge; however, the supervised datasets by manual annotation are only maintained at the level of 100,000, which is far lower than the data scale in natural language processing, such as text classification and entity recognition. For deep neural network models, including convolutional neural network (CNN) [17], recurrent neural network (RNN) [18], and long short-term memory network (LSTM) models [14,19], their accuracy and generalizability are difficult to guarantee with small amounts of supervised data. Even for a small amount of supervised data, introducing and combining unsupervised learning strategies for extracting drug–drug interaction also needs to be studied.

To address the above problems, we propose a novel method called Entity-Mask Symbol-Insert BERT (EMSI-BERT) that incorporates the asymmetrical Entity-Mask strategy into the pre-training and the Symbol-Insert structure into the fine-tuning of BERT to realize end-to-end DDI extraction. The framework of the proposed method is shown in Figure 1.

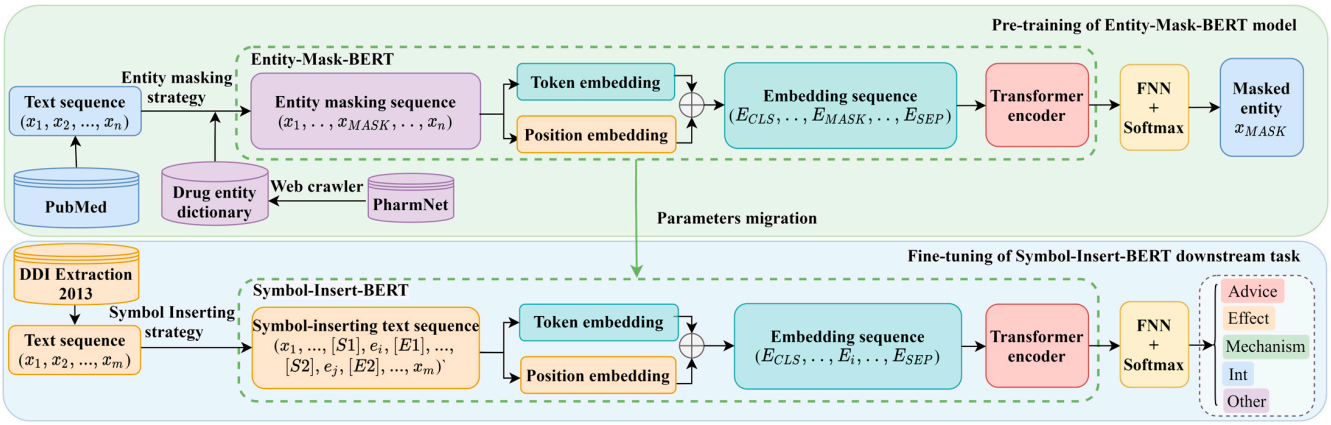

Figure 1. Framework of the asymmetrical Entity-Mask strategy and the Symbol-Insert structure for drug–drug interaction extraction based on BERT (FNN represents Feed-forward Neural Network).

Firstly, after using a web crawler to collect massive drug text data, we construct a drug entity dictionary and explore drug entities with string matching algorithms. Secondly, to address the lack of entity orientation in the random masking of the original pre-training BERT, we propose an asymmetrical Entity-Mask strategy to improve the expression of co-occurring entity information using the drug entity dictionary. Finally, we design the Symbol-Insert structure to address DDI classifications without destroying the pre-training BERT. Four symbols (S1, E1, S2, and E2) are inserted in the same input sequence according to the position of entity pairs to ensure that the same input sequence has different forms in the input layer. More specifically, this paper makes the following contributions:

1. The pre-training strategy of random masking is improved. In the pre-training BERT, an asymmetrical Entity-Mask strategy is proposed to compensate for the lack of entity orientation in the random masking strategy. Based on prior knowledge, the mask probability of drug entities is increased to better retain entities' co-occurrence infor-

mation. Ablation experiments confirm that the pre-training BERT with asymmetrical Entity-Mask strategy effectively improves the effect of downstream DDI classification.

2. The fine-tuning structure to adapt to downstream tasks is investigated. In the fine-tuning BERT, a Symbol-Insert structure is proposed to preserve most of the structural information of the pre-training BERT and overcome the problem of different entity combinations sharing the same input sequence. The same input sequence is given different forms in the input layer by adding four symbols to the entity combinations, thereby allowing DDI extraction without destroying the structure of pre-training BERT. The experimental results show that the proposed structure can be adapted to the DDI extraction task effectively.
3. The migration scheme of combining pre-training and fine-tuning is proposed. An EMSI-BERT method, which incorporates the asymmetrical Entity-Mask strategy into the pre-training and the Symbol-Insert structure into the fine-tuning of BERT, is proposed to realize DDI extraction with few labeled data. Compared with related methods, the proposed EMSI-BERT method is insensitive to data preprocessing and demonstrates comprehensive improvement in the two-classification task of DDI detection and the multi-classification task of DDI extraction, including Advise, Effect, Mechanism, and Int.

After Section 2 discusses related work on DDI extraction, Section 3 introduces the pre-training BERT. The proposed EMSI-BERT for DDI extraction is detailed in Section 4. Section 5 then presents the results and discussion of the proposed method, and compares them with those of other related methods. Finally, Section 6 summarizes this study.

2. Related Work

Researchers have recently developed many DDI extraction methods, which fall primarily into two categories: rule-based and statistical machine learning-based methods [1,2], as shown in Table 1. Rule-based methods [20–23] mainly use predefined rules and template matching techniques to extract the relationship of drug entities in the sentences. These methods with customized templates are highly accurate with a relatively low recall rate because templates are limited. Statistical machine learning methods, which include traditional machine learning-based and deep learning-based methods, regarded DDI extraction as a classification problem, i.e., whether the two entities in the same sentence interact with each other is judged. The traditional machine learning-based methods need to define many features, including an n-gram, syntactic tree, dependency tree, and other information [24–30]. The advantages of these methods offer a balance between accuracy and recall through various classification models, including the maximum entropy model and support vector machine (SVM). Thus, the traditional machine learning-based methods have better effects and advantages in early DDI classification research [5,24–26].

The traditional machine learning-based methods require manual feature extraction and cascading of different features. Unlike these, deep learning-based DDI classification methods achieve DDI extraction by directly inputting the text description of the drug–drug relation. This end-to-end style reduces the complexity of manual feature extraction and avoids the error accumulation of cascading external models (such as syntactic analysis models). The relevant literature shows that the experimental effects of CNN-, RNN-, and BERT-based methods surpass traditional machine learning-based methods [6,31,32]. This improvement is mainly attributed to the following reasons: (1) Words are expressed as dense vectors instead of sparse one-hot variables. Representing words as dense vectors, such as word2vec [33], allows previously incomparable features to be effectively measured in the vector space. For instance, in DDI extraction, both "rely" in "rely on" and "depend" in "depend on" indicate dependence. This connection is lost in one-hot representation but can be calculated in the vector space through the dot product between word vectors. (2) Deep neural network-based methods can fit in a high dimensional space by constructing various hyper-plane spaces through massive parameters. Representative works, including CNN-, RNN-, and BERT-based methods, have shown promising results in DDI tasks [18,34].

In CNN-based methods [6,17,35–38], words in an input sentence containing entity pairs are first converted into word vector representations. Next, the CNN structure is used to capture the n-gram related information and acquire the semantic representation of the input sentence. The relationship between entity pairs is then classified based on semantic representation. In RNN-based methods [14,17–19,31,34,39–44], a sentence is directly subjected to sequence modeling to obtain its semantic representation between entity pairs. In BERT-based methods [15,45], BERT is regarded as a feature extractor for obtaining entity-level representations, and entity pair classification and relation extraction are then conducted through downstream domain tasks [16]. However, most existing methods require a large amount of external information to achieve a better understanding [32,46]. On the whole, insufficient domain-labeled data, weakly expressed co-occurring entities, and poor adaptation of downstream tasks can be factors limiting the classification ability of a pre-training BERT model.

Table 1. Comparison with the advantages and disadvantages of different methods (Rule-based methods, Traditional machine, and Deep learning-based methods).

Taxonomy	Method	Advantages	Disadvantages
Rule-based methods	Bunescu et al. [20], Fundel et al. [21], Segura-Bedmar et al. [22], An et al. [23]	These methods with customized templates are highly accurate for DDI extraction.	(1) The design of patterns or rules is sophisticated; (2) These methods suffer from low recall because of limited templates.
Traditional machine learning-based methods	Cui et al. [24], Segura-Bedmar et al. [25], Kim et al. [26], FBK-irst [27], WBI [28], UTurku [29], RBF-Linear [30]	These methods offer a balance between accuracy and recall through various classification models.	(1) The design of hand-crafted features is sophisticated; (2) The cascading strategy of different features requires is elaborate designed.
Deep learning-based methods	CNN [10,35–38], RNN [9–14,14–17,31,34,40–44], BERT [15,45]	These end-to-end methods reduce the complexity of manual feature extraction and avoid the error accumulation of cascading external models.	(1) These methods require a large amount of external information to achieve a better understanding for DDIs extraction; (2) These methods are poorly suited to co-occurring entities expression and adaptation of downstream tasks.

3. Materials and Methods

In recent years, pre-training models have received increasing attention from researchers, such as the ELMO (an RNN-based structure) [47] and BERT (a Transformer-based structure) [45,48]. BERT, which is a pre-training model based on bidirectional encoder representations from Transformers, has been proved to be effective for improving many natural language processing tasks. In order to better adapt downstream tasks and reduce the consumption of computing power and time, BERT only embeds the encoder structure of Transformers for two subtasks: masking language prediction and sentence pair classification.

As shown in Figure 2, the pre-training BERT model consists of three layers: input, encoding, and output. For the masking language prediction subtask, in the input layer, [CLS] and [SEP] symbols are firstly added at the beginning and end of an input sentence. Then, the words in the input sentence are randomly masked, that is, the randomly selected words are covered with the [MASK] symbol. Researchers generally use three masking strategies to simulate the data distribution in real-world scenarios: replacing 80% of the masked words with [MASK], replacing 10% of the masked words with random words, and

keeping 10% of the masked words unchanged [48]. Figure 2 shows the masking strategy, replacing the words "the" and "of" with [MASK]. Finally, the embedding sequence, which can be obtained by token embedding, segment embedding, and position embedding from the masking sentence, is fed into the encoding layer based on the Transformer blocks. Generally, multi-layers of Transformer blocks are stacked to improve the expression ability of the pre-training model. In the output layer, the masked words in the input layer are predicted according to contextual information extracted from the encoding layer. The masked language prediction task in Figure 2 involves predicting the masking words "the" and "of" in the output layer. The masking language prediction subtask reveals that the pre-training process of BERT requires no labeled data. Hence, a large batch of unsupervised text data can be fully used. In essence, the BERT mask prediction task is considered a mask-based reconstruction with added noise. Through this reconstruction in the input layer, the pre-training model can learn the internal connection between the masked words and the remaining words, which is conducive to representing the general semantic information of the input sentence. Sentence pair classification is another BERT subtask, and it mainly concerns whether a pair of sentences appears in the same document. Considering it has a small correlation with DDI extraction and consistent with [16,49], this paper will also discard the sentence pair classification subtask.

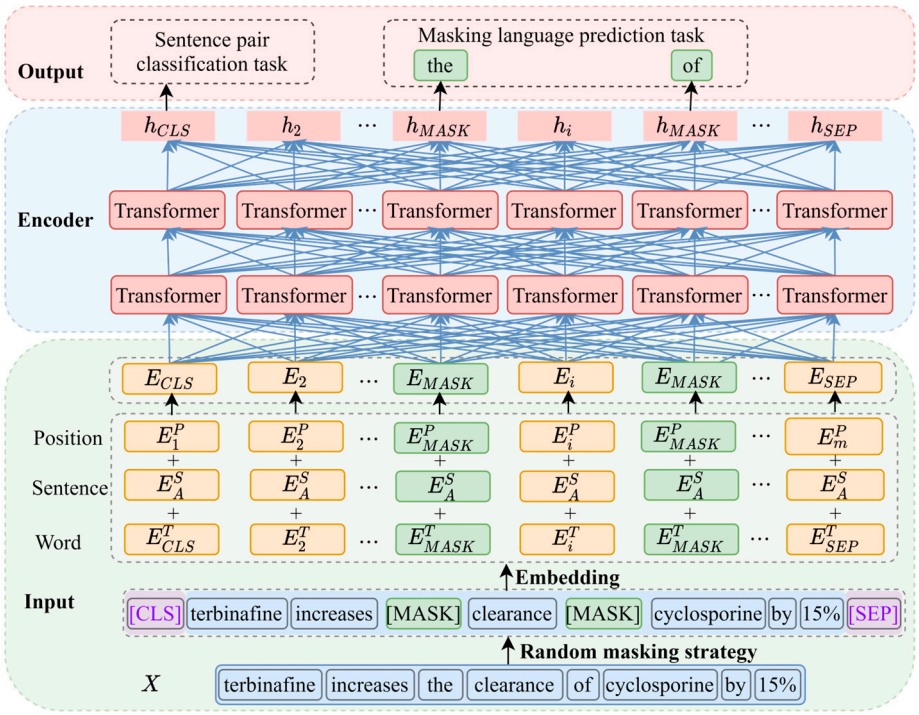

Figure 2. Framework of pre-training BERT (X represents Input sequence; E_i^T, E_A^S, and E_i^P represent Token embedding, Segment embedding, and Position embedding, respectively; E_i and h_i represent Embedding vector and Latent semantics of the i-th word, respectively).

4. EMSI-BERT for DDI Extraction

To address the shortcomings of a lack of entity propensity in the random masking strategy of the pre-training BERT, we propose an asymmetrical Entity-Mask strategy to improve the masking language model combining with drug entity recognition. Then, a Symbol-Insert structure for fine-tuning BERT is designed based on the pre-training BERT with an asymmetrical Entity-Mask strategy.

4.1. An Asymmetrical Entity-Mask Strategy for Pre-Training BERT

As shown from the training strategy of the mask language prediction task in Figure 2, BERT masks words in the input sequence randomly. Due to the absence of prior knowledge,

such a propensity-free random strategy masks words that contain less information, as well as words of less relevance and attention to the relation extraction task. In Figure 2, the words "the" and "of" have a relatively low amount of information but are still masked and predicted. These masked input sequences carrying redundant information not only constrain the representation capability of the encoding layer, but also reduce the training difficulty, causing the model to encode shallow information. To eliminate random masking, we propose an Entity-Mask-BERT with an asymmetrical Entity-Mask strategy, presented in Figure 3.

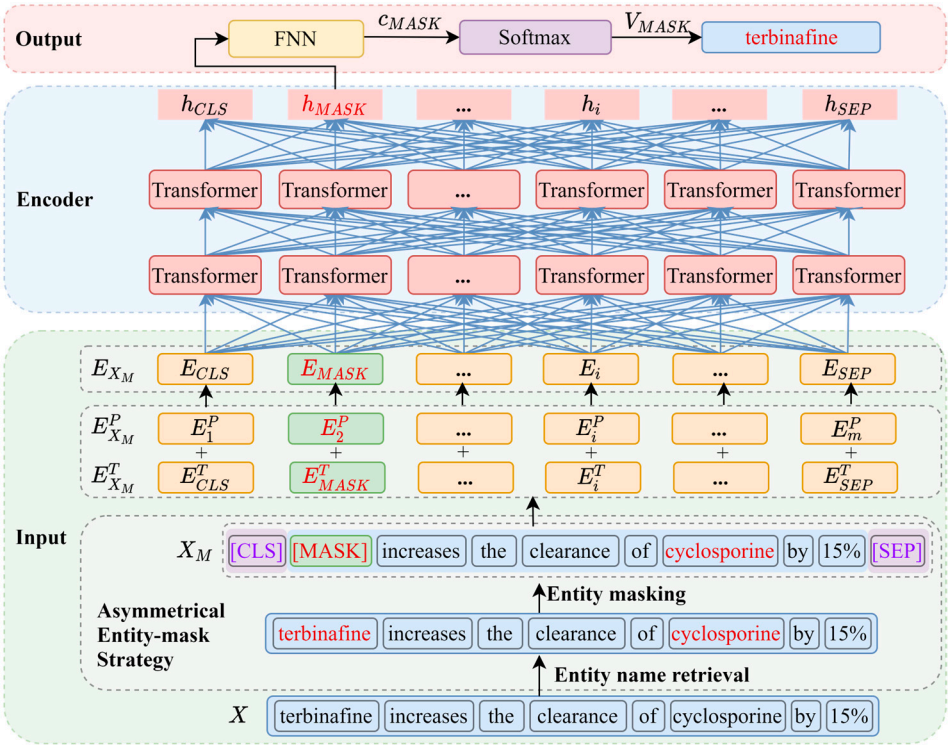

Figure 3. Asymmetrical Entity-Mask strategy for pre-training BERT (FNN represents Feed-forward Neural Network; X and X_M represent Input sequence and Masking sequence, respectively; $E^T_{X_M}$, $E^P_{X_M}$, and E_{X_M} represent Token embedding sequence, Position embedding sequence, and Embedding representation of X_M, respectively; c_{MASK} represents unnormalized category probability vector; V_{MASK} represents normalized category probability vector).

4.1.1. Entity-Mask-BERT Model Construction

For the drug relationship description sequence $X = (x_1, x_2, \cdots, x_n)$ with the input length of n, drug entity detection is firstly realized by string matching based on the self-built drug entity dictionary during the sequence masking stage in the input layer. If the number of drug entities is fewer than 2 in the input sequence, the completely random mask strategy will be followed. Otherwise, the asymmetrical Entity-Mask strategy is addressing drug entity masking. An example is shown in Table 2.

Table 2. Asymmetrical Entity-Mask strategy for the sentence "terbinafine increases the clearance of cyclosporine by 15%".

Drug Entity A	Drug Entity B	Example
Replaced with [MASK]	Reserved	[MASK] increases the clearance of cyclosporine by 15%
Reserved	Replaced with [MASK]	Terbinafine increases the clearance of [MASK] by 15%

In Table 2, the words "terbinafine" and "cyclosporine" in the sentence "terbinafine increases the clearance of cyclosporine by 15%" are recognized as the drug entity names. Instead of randomly selecting a word from the sentence, we firstly replace "terbinafine"

with [MASK] and reserve "cyclosporine" to obtain a new entity-making sequence "[MASK] increases the clearance of cyclosporine by 15%" for pre-training. Secondly, we reserve "terbinafine" and replace "cyclosporine" with [MASK] to obtain another entity-making sequence "terbinafine increases the clearance of [MASK] by 15%" for pre-training. Table 2 shows that two entity-making sequences are generated for the sequence containing two drug entities based on the asymmetrical Entity-Mask strategy. Similarly, C_N^2 entity-making sequences will be generated for pre-training when the input sequence contains N drug entities. The proposed entity masking strategy is mainly inspired by the concept of distant-supervised relation extraction [50,51]. When multiple entities appear in the input sequence, the proposed strategy masks one entity and utilizes other drug entities to predict the currently masked entity. This strategy can achieve unsupervised high frequency drug–drug relation modeling and complete pre-training tasks with a large amount of unlabeled drug mechanism data. Except for the entity words, the remaining words refer to the completely random mask strategy.

Suppose the input sequence after the drug entity masking is $X_M = (x_1, \cdots, [MASK], \cdots, x_i \cdots, x_n)$. In addition, we add the [CLS] and [SEP] symbols to the beginning and end of X_M, respectively. To obtain the word token embedding representation $E^T_{X_M}$, words in the input sequence are then divided into a limited set of common word units according to the WordPiece Embedding technique [15,45,46]:

$$E^T_{X_M} = (E^T_{CLS}, \cdots, E^T_{MASK}, \cdots, E^T_i, \cdots, E^T_{SEP}) = \text{WordPiece}([CLS, X_M, SEP]) \quad (1)$$

where $E^T_{X_M} \in R^{d \times m}$ is the word token embedding representation of the masking sequence X_M; $d = 768$ and m ($m \geq n + 2$) represent the dimension of word embedding and the number of common word units divided from the masking sequence, respectively. However, the word token embedding only reflects the word information in the sequence and lacks the sequential relation between words. Experience has shown that the same word has completely different meanings depending on where it appears in a sequence. To reflect the position information of different words in the sequence, we adopt absolute position coding to realize the position embedding of the sequence.

$$E^P_{X_M} = (E^P_{CLS}, \cdots, E^P_{MASK}, \cdots, E^P_i, \cdots, E^P_{SEP}) \quad (2)$$

where $E^P_{X_M} \in R^{d \times m}$ is the position embedding representation of the masking sequence X_M. After word token embedding and position embedding, the final embedding representation of X_M can be expressed as

$$E_{X_M} = (E_{CLS}, E_2, \cdots, E_{MASK}, \cdots, E_{SEP}) = E^T_{X_M} + E^P_{X_M} \quad (3)$$

Different from Figure 2, the sentence pair classification task is discarded in Entity-Mask-BERT. Therefore, the sentence embedding is not required in the input layer.

After obtaining the embedding sequence E_{X_M}, we then extract the global semantic feature of each word by inputting E_{X_M} into a 12-layer transformer encoder.

$$H = (h_{CLS}, \cdots, h_{MASK}, \cdots, h_i, \cdots, h_{SEP}) = Encoder(E_{CLS}, \cdots, E_{MASK}, \cdots E_i, \cdots, E_{SEP}) \quad (4)$$

where h_i is the global semantics of the embedding vector E_i and $Encoder(\cdot)$ represents a 12-layer transformer blocks. As can be seen from the model structure, it contains the context information of the sequence and enhances its own characteristics, allowing the model to capture information E_i based on all contexts.

Finally, to predict the masked entity x_{MASK} in the output layer, the global semantic feature h_{MASK} is fed to a full connection layer:

$$c_{MASK} = FNN(h_{MASK}) = W_1 h_{MASK} + b_1 \quad (5)$$

where $FNN(\cdot)$ represents feed-forward neural network. $W_1 \in R^{N\times d}$ and $b_1 \in R^N$ is weight matrix and bias, respectively. $c_{MASK} = (c^1, \cdots, c^j, \cdots, c^N) \in R^N$ represents the unnormalized category probability. N represents the number of piece words in the dictionary. Moreover, a Softmax function is used to obtain the normalized category probability:

$$v^j = \frac{e^{c^j}}{\sum_{j=1}^{N} e^{c^j}} \tag{6}$$

where $V_{MASK} = (v^1, \cdots, v^j, \cdots, v^N) \in R^N$ is the normalized category probability vector. v^j $(1 \le j \le N)$ represents the normalized probability that the masking entity x_{MASK} belongs to the j-th word.

4.1.2. Pre-Training of Entity-Mask-BERT

To train Entity-Mask-BERT, we regard the mask-entity prediction task as a multi-classification containing N types of words. Let the masking words in the input layer be $S = \{x_i\}_{i=1}^{s}$ and their category probability obtained by Entity-Mask-BERT be $\left\{V_i = (v_i^1, \cdots, v_i^j, \cdots, v_i^N) \in R^N\right\}_{i=1}^{s}$, the objective function is constructed as follows:

$$L_1(\theta_B, \theta_M) = -\sum_{i=1}^{m}\sum_{j=1}^{N} u_i^j \log v_i^j \tag{7}$$

where m is the number of masked drug entities; θ_B and θ_M are the parameters of encoder layer and output layer in Entity-Mask-BERT, respectively; $\log(\cdot)$ is a logarithmic function; and $u_i = (u_i^1, \cdots, u_i^j, \cdots, u_i^N)$ and $v_i = (v_i^1, \cdots, v_i^j, \cdots, v_i^N)$ represent the true label vector and the predicted label vector of the masking entity x_i, respectively. Related hyperparameters of Entity-Mask-BERT are shown in Table 3.

Table 3. Related hyperparameters of Entity-Mask-BERT.

Hyperparameter	Value
Optimizer	Adam
Learning rate	1×10^{-5}
Warm-up rate	0.1
Batch-size	256
Sentence length m	Dynamic padding
Dimension of word embedding d	768
Number of Transformer blocks	12

4.2. A Symbol-Insert Structure for Fine-Tuning BERT

In addition to the entity masking strategy introduced in the pre-training BERT, DDI extraction quality is also closely related to constructing the downstream domain task. If the input sequence contains N entities, C_N^2 entity combinations must be ascertained. How to distinguish C_N^2 combinations and to classify the relation of each combination in the same input sequence is a pivotal issue to be solved in this paper. Specifically, in the input sequence "Grepafloxacin, like other quinolones, may inhibit the metabolism of caffeine and theobromine", the words "Grepafloxacin", "caffeine", and "theobromine" represent three drug entities. Mathematically, the relations among three drug entity combinations [Grepafloxacin, caffeine], [Grepafloxacin, theobromine], and [caffeine, theobromine] need to be identified in the downstream domain task. However, the same input sequence is shared among these entity combinations. To realize combination discrimination, we design a Symbol-Insert structure to extract entity relations among different drugs in the same input sequence by introducing four symbols.

4.2.1. Symbol-Insert-BERT Model Construction

To distinguish different entity combinations in the same input sentence and not destroy the structure of the original BERT, we introduce four novel positional symbols S1, E1, S2, and E2. For the input sequence above, to validate the relations between entity combinations [Grepafloxacin, caffeine], we insert the symbols S1 and E1 before and after "Grepafloxacin" to mark the position of the first entity, and then insert the symbols S2 and E2 before and after "caffeine" to mark the position of the second entity. Likewise, these operations are performed in the entity combinations [Grepafloxacin, theobromine] and [caffeine, theobromine]. The different symbol inserting results for three entity combinations are shown in Figure 4.

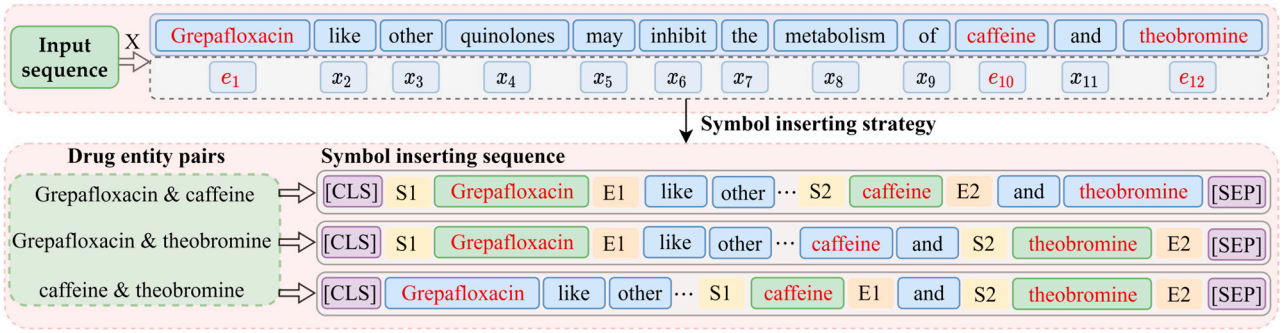

Figure 4. Symbol inserting sequence for different entity combinations (S1 and E1 represent the position of the first entity; S2 and E2 represent the position of the second entity).

According to Figure 4, an input sentence containing N entities generates C_N^2 kinds of symbol inserting sequences by introducing four positional symbols, thus facilitating different representations of the same input sequence in the input layer. The specific construction form is presented in Figure 5. In Figure 5, the parameters of Entity-Mask-BERT are utilized for the initialization of fine-tuning BERT with Symbol-Insert structure. The constructed Symbol-Insert structure has the following two advantages: firstly, the positional symbols are only added in the input layer, avoiding changing the internal Transformer structure in the encoding layer or the overall framework, and ensuring that the pre-training BERT with the asymmetrical Entity-Mask strategy can favorably be transferred to the downstream task of DDI classification; secondly, the insertion of the position symbols only alters the relative position relation among the original input words. It is well known that the attention mechanism of the Transformer is insensitive to positional alternation, which permits a relative concordance between the improved pre-training strategy and the fine-tuning scheme with the Symbol-Insert structure.

4.2.2. Fine-Tuning of Symbol-Insert-BERT

The transformer blocks are the fundamental encoding unit in the fine-tuning BERT with the Symbol-Insert structure, and its parameters are all initialized by Entity-Mask-BERT. Consistent with other BERT classification tasks, the output of the [CLS] node is also adopted as a relational representation of drug entity pairs and used to predict the result of DDI classification. During the model training stage, the output representation of the [CLS] node is assumed to be $r \in R^k$. Firstly, the full connection layer is adopted to map the output from the representation space to the category probability space:

$$o = FNN(r) = W_2 r + b_2 \tag{8}$$

where $O = (o_1, \cdots, o_i, \cdots, o_c) \in R^C$ represents the unnormalized probability of each category; $W_2 \in R^{C \times k}$ and $b_2 \in R^C$ are transfer matrix and bias, respectively; k and C represent the output vector dimension and the number of categories of drug entity relations, respectively.

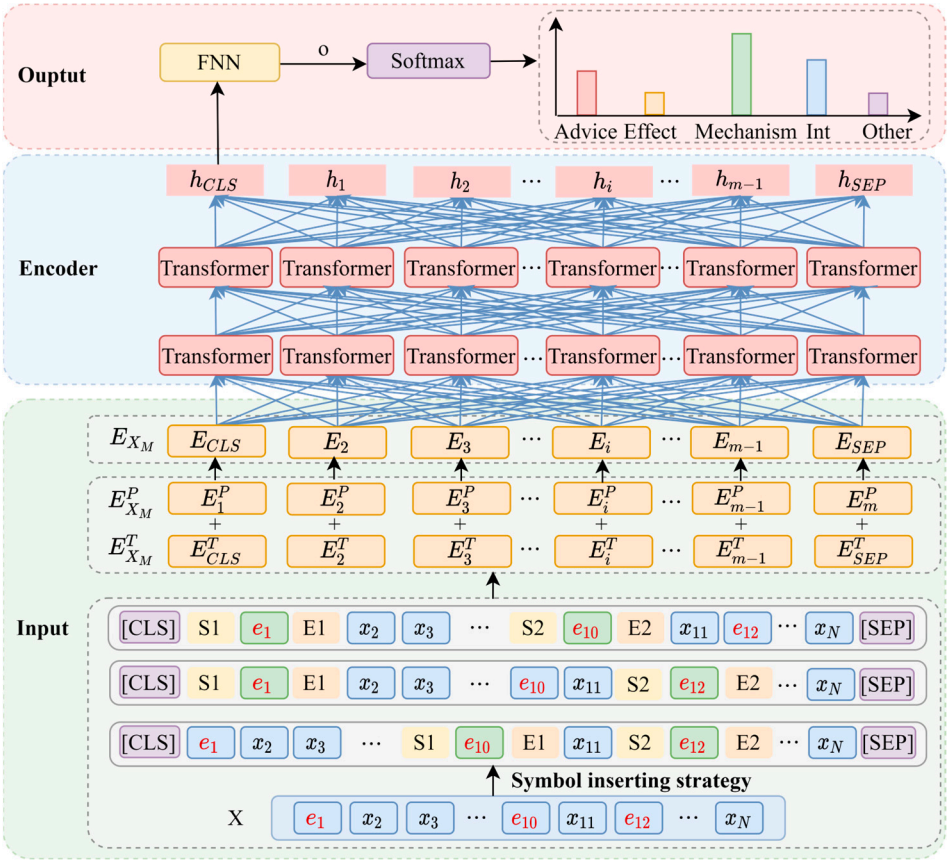

Figure 5. Construction of the Symbol-Insert-BERT structure (FNN represents Feed-forward Neural Network; X represents Input sequence, where e_1, e_{10}, and e_{12} denote different drug entities, and x_2, x_3, x_{11}, and x_N denote non-drug entities; $E^T_{X_M}$, $E^P_{X_M}$, and E_{X_M} represent Token embedding sequence, Position embedding sequence, and Embedding representation, respectively; O represents the unnormalized category probability).

Secondly, the unnormalized probability O is transformed into the normalized probability S by the Softmax function:

$$S_j^{l,p} = p\left(c_j \middle| X_p^l\right) = \frac{e^{o_j}}{\sum_{i=1}^{C} e^{o_i}} \tag{9}$$

where X_p^l represents the pth drug entity pair in the lth input sentence (sample), and $S_j^{l,p}$ denotes the probability that X_p^l belongs to the category c_j.

Finally, Symbol-Insert-BERT is trained by the following optimization objective:

$$L_2 = \sum_{l=1}^{L} \sum_{p=1}^{n_l} \sum_{j=1}^{C} -y_j^{l,p} \log S_j^{l,p} \tag{10}$$

where L represents the total number of samples in the training set; n_l indicates the number of drug entity pairs in the lth sample; $y_j^{l,p}$ and $S_j^{l,p}$ denote the actual and predicted values of the pth drug entity pair in the lth sample, respectively.

In the fine-tuning of Symbol-Insert-BERT, the output vector dimension is $k = 768$ and the number of categories is $C = 5$, the hyperparameter of batch-size is set to 16, while other hyperparameters are consistent with Entity-Mask-BERT. In particular, if the word vectors for the position symbols S1, E1, S2 and E2 are added as additional parameters, the size of the word vector dictionary in the pre-trained BERT model will inevitably be disrupts. Since

BERT reserves part of the word vector units for the new symbols, we map the positional symbols to the [unseen] symbol in the BERT's vocabulary without increasing the BERT vocabulary size. The proposed fine-tuning Symbol-Insert with pre-training parameters will converge in about 20 iterations. We provided the implementation details and source code of the proposed method on the GitHub repository [52].

5. Results and Discussion

5.1. Biomedical Corpus for Pre-Training BERT with Asymmetrical Entity-Mask Strategy

The pre-training corpus for Google's BERT model is mainly from Wikipedia. Applying the BERT model to DDI extraction, we use the pre-training model of Google's BERT as a basis and select the abstract data in the PubMed dataset as the corpus for pre-training Entity-Mask-BERT. PubMed dataset [53], which contains medical and health literature published from 1966, is a free biomedical network resource. Currently, this database is updated weekly, and 35 million citations for biomedical literature are collected as biomedical corpus for pre-training Entity-Mask-BERT. In addition, drug entity detection is needed for the asymmetrical Entity-Mask strategy. Hence, a large number of drug entity names are collected via web crawlers at PharmNet [54] for the drug entity dictionary construction. The self-built dictionary, including 100,000 drug entity names, is addressed to drug entity recognition by string matching. In Entity-Mask-BERT, when the word in one sequence matches the drug entity in the dictionary, we regard it as drug entities and adopt the improved asymmetrical Entity-Mask strategy to obtain the entity-making sequence for masking language prediction. After preparing pre-training data and the drug entity dictionary, the pre-training BERT with asymmetrical Entity-Mask strategy is performed.

5.2. Domain-Labeled Dataset for Fine-Tuning BERT with Symbol-Insert Structure

The DDI-Extraction 2013 dataset provides a more extensive annotated corpus for DDI extraction. The training and test sets of DDI-Extraction 2013 are composed of two main aspects, DrugBank and MedLine. The original data statistics of this dataset is shown in Table 4. In the dataset, drug entity pairs with five types of DDI, including Advice, Effect, Mechanism, Int, or Other (no relation between two entities) [19] is labeled for the five-label multi-classification task. Each label is briefly described as follows.

(1) Advice: describes the relevant opinion about the simultaneous use of two drugs, i.e., interaction may be expected, and UROXATRAL should not be used in combination with other alpha-blockers;
(2) Effect: describes the interaction of drug effects, i.e., methionine may protect against the ototoxic effects of gentamicin;
(3) Mechanism: describes the pharmacokinetic mechanism, i.e., Grepafloxacin, like other quinolones, may inhibit the metabolism of caffeine and theobromine;
(4) Int: describes the DDI without any information, i.e., the interaction of omeprazole and ketoconazole has been established;
(5) Other: describes co-occurrence but no relation between two entities, i.e., concomitantly given thiazide diuretics did not interfere with the absorption of a tablet of digoxin.

Table 4. Data statistics of the original DDI-Extraction 2013 dataset.

Relation	Train			Test		
	DrugBank	**MedLine**	**Overall**	**DrugBank**	**MedLine**	**Overall**
Advice	818	8	826	214	7	221
Effect	1535	152	1687	298	62	360
Mechanism	1257	62	1319	278	24	302
Int	178	10	188	94	2	96
Other	22,118	1547	23,665	4367	345	4712

As can be seen from Table 4, the number of instances belonging to different types is extremely unbalanced, which makes it difficult to classify the drug relationship with fewer instances. For example, there are more than 100 times as many instances of "Other" as instances of "Int" in the training set. To eliminate the effect of massive "Other" samples and enhance the effect of the model, we filter out partial "Other" sample pairs in the data processing stage via the text preprocessing approach proposed by Quan [36]. The pre-processed DDI-Extraction 2013 dataset is shown in Table 5. Table 5 illustrates that the text preprocessing approach allows the reduction of "Other" samples from 22,118 to 14,445 in the training set and the reduction of "Other" samples from 4367 to 2819 in the test set. The preprocessing approach reduces the unbalance of the classification while saving training time. Finally, the training and testing samples in Table 5 are used as the training set and testing set for Symbol-Insert BERT in fine-tuning stage, respectively.

Table 5. Data statistics of DrugBank and MedLine after preprocessing.

Relation	Train			Test		
	DrugBank	**MedLine**	**Overall**	**DrugBank**	**MedLine**	**Overall**
Advice	815	7	822	214	7	221
Effect	1517	152	1669	298	62	360
Mechanism	1257	62	1319	278	21	299
Int	178	10	188	94	2	96
Other	14,445	1179	15,624	2819	243	3062

5.3. Experimental Results and Analysis

To evaluate the performance of different methods, precision (P), recall (R), and F1 score are regarded as the evaluation indicators of the DDI classification task. The indicator values for the jth ($1 \leq j \leq C$) drug relationship are calculated as follows:

$$P_j = \frac{TP_j}{TP_j + FP_j}, R_j = \frac{TP_j}{TP_j + FN_j}, F1_j = \frac{2P_jR_j}{P_j + R_j} \ (1 \leq j \leq C) \tag{11}$$

where TP_j represents the number of both the predicted relations and the true relations belonging to the j-th category. FP_j represents the number of predicted relations that does belong to the j-th category but the true relation does not. FN_j represents the number of predicted relations that do not belong to the j-th category but the true relation does. $TP_j + FP_j$ and $TP_j + FN_j$ represent the total number of the predicted relation and the true relation belonging to the j-th relation, respectively. In addition, considering that DDI extraction is a multi-class classification, the micro-averaged F1-score $F1_{micro}$ is calculated to evaluate the overall classification performance of different methods:

$$P_{micro} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C} TP_j + \sum_{j=1}^{C} FP_j}, R_{micro} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C} TP_j + \sum_{j=1}^{C} FN_j}, F1_{micro} = \frac{2P_{micro}R_{micro}}{P_{micro} + R_{micro}} \tag{12}$$

To ensure the stability of statistical indicators and illustrate the convergence of the proposed method, 10-fold cross-validation is used to train the Entity-Mask-BERT and Symbol-Insert-BERT. Meanwhile, to make the experimental results more reliable, 10-fold cross-validation is conducted 10 times, and the average is used as the final result of indicators.

5.3.1. Performance Evaluation of the Proposed Method

To illustrate the performance of the proposed method, we compare the effect of Symbol-Insert-BERT initialized with Basic-BERT and Entity-Mask-BERT on DDI extraction. The Basic-BERT model and parameters can be obtained from the Hugging Face website [55].

The proposed Entity-Mask-BERT and Basic-BERT both adopt a 12-layer-768 structure. The experimental results of two initialization strategies are shown in Table 6, and the confusion matrix for two initialization strategies are shown in Tables 7 and 8, respectively.

Table 6. Evaluation of DDI classification performance with Basic-BERT initialization and Entity-Mask-BERT initialization.

Relation	Basic-BERT Initialization + Symbol-Insert-BERT			Entity-Mask-BERT Initialization + Symbol-Insert-BERT (EMSI-BERT)		
	P	R	F1-Score	P	R	F1-Score
Advice	87.85	85.84	86.83	85.52	88.20	86.86
Effect	77.89	81.18	79.50	79.56	82.2	80.77
Mechanism	82.97	76.84	79.79	88.46	84.89	86.64
Int	66.17	46.8	54.87	72.13	45.83	56.05
Other	80.83	77.50	79.13	83.22	80.74	81.96

Table 7. Confusion matrix of DDI classification with Basic-BERT initialization.

Relation	Advice	Effect	Mechanism	Int	Other
Advice	85.8	0.4	0	0.9	12.7
Effect	1.6	81.1	2.8	0.2	14
Mechanism	2	1.3	76.8	3.3	16.4
Int	0	40.6	2	46.8	10.4
Other	0.4	1.3	1.2	0.3	96.6

Table 8. Confusion matrix of DDI classification with Entity-Mask-BERT initialization.

Relation	Advice	Effect	Mechanism	Int	Other
Advice	88.2	0.4	0	0.9	10.4
Effect	3.3	82	0.8	0	13.7
Mechanism	2.3	1.6	84.8	0	11
Int	0	38.5	2	45.8	13.5
Other	0.2	0.6	0.5	0.3	98.1

Table 6 shows that three statistics of EMSI-BERT achieve better results compared with the Basic-BERT initialization. The overall P, R, and F1-score are increased by 2.39, 3.24, and 2.83, respectively. For individual DDI classification, EMSI-BERT also obtains a high F1-score, which illustrates the importance and effectiveness of the asymmetrical Entity-Mask strategy. Specifically, three indicators of "Effect" and "Mechanism" are significantly improved. The confusion matrix for two initialization strategies in Tables 7 and 8 show that main errors focus on two areas: (1) four relation categories (Advice, Effect, Int, and Mechanism) are misclassified as "Other". The main reason for this phenomenon is that the data ratio of four relation categories is still much lower than that of "Other" category. Even though the data is imbalanced, the misclassification of EMSI-BERT is well suppressed compared with Basic-BERT initialization. (2) Three indicators of "Int" are comparatively low, "Int" and "Effect" misclassify each other. The main reason for this phenomenon is that the number of "Int" instances is too small, and some instances of "Int" and "Effect" have similar semantics. Moreover, our statistics indicates that about 10% of the drug pairs labeled as "Int" are also labeled as "Effect" in the training set.

In addition, we further discuss the impact of epochs on the performance of EMSI-BERT, as show in Figure 6. Figure 6 shows that the micro-averaged F1-score is an upward trend with the increase of the number of epochs. When the epoch reaches 16 times, the micro-averaged F1-score attains a peak and tends to be stable. This indicates that the proposed EMSI-BERT has good convergence. Furthermore, the output feature h_{CLS} of the final (12-th)

layer in Transformer blocks is fed to the output layer for DDI extraction in EMSI-BERT. To illustrate the impact of different output features of Transformer layers on the micro-averaged F1-score, the output feature of the 6th, 9th and 12th layer is fed to the output layer, respectively. Figure 7 shows that the features learned by Transformer layers are different and taking the deepest semantic features as input can obtain the best performance.

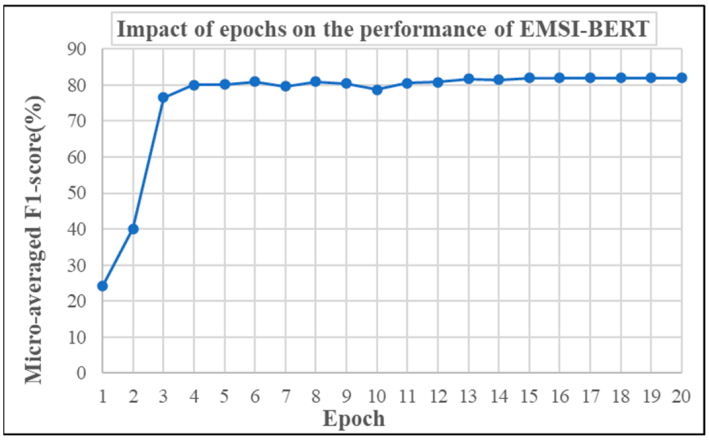

Figure 6. Impact of epochs on the performance of EMSI-BERT.

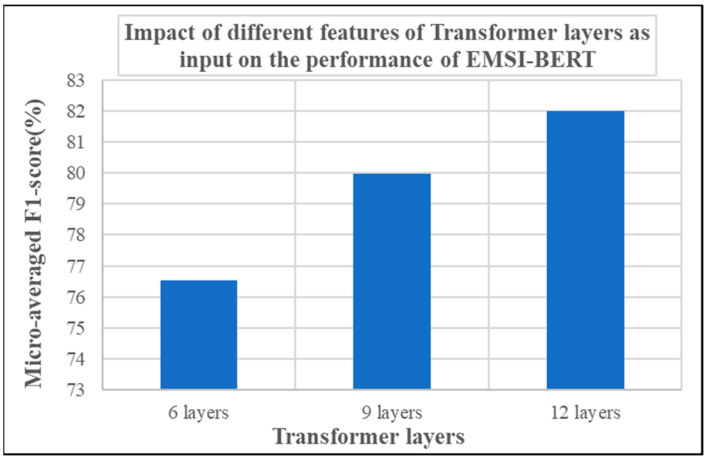

Figure 7. Impact of different features of Transformer layers as input on the performance of EMSI-BERT.

5.3.2. Comparison of DDI Classification with Related Methods

Using the DDI-Extraction 2013 dataset, this section compares the effectiveness of the proposed EMSI-BERT method for DDI classification with other related work. In addition to considering four DDI multi-classification tasks, including Advice, Effect, Mechanism, and Int, we compare the effectiveness of the two-classification task of drug relationship detection, that is, Dec in Table 9, and Table 10 is introduced to describe the presence or absence of DDI without distinguishing the kind of relationship.

Table 9. Comparison of DDI classification with traditional statistical machine learning-based methods.

Method	Advice	Effect	Mechanism	Int	Dec	Micro-Averaged F1-Score
Kim et al. [26]	72.5	66.2	69.3	48.3	77.5	67.0
FBK-irst [27]	69.2	62.8	67.9	54.0	80.0	65.1
WBI [28]	63.2	61.0	61.8	51.0	75.9	60.9
UTurku [29]	63.0	60.0	58.2	50.7	69.6	59.4
RBF-Linear [30]	77.4	69.6	73.6	52.4	81.5	71.1
EMSI-BERT	86.8	80.7	86.6	56.0	88.0	82.0

Table 10. Comparison of DDI classification with deep learning-based methods (CNN, RNN and BERT).

Model	Method	Advice	Effect	Mechanism	Int	Dec	Micro-Averaged F1-Score
CNN	CNN [35]	77.7	69.3	70.2	46.3	-	69.8
	SCNN [37]	- *	-	-	-	77.2	68.4
	MCNN [36]	78.0	68.2	72.2	51.0	79.0	70.2
	RHCNN [38]	80.5	73.5	78.3	58.9	-	75.5
	AGCN [10]	86.2	74.2	78.7	52.6	-	76.9
RNN	Hierarchical RNN [18]	80.3	71.8	74.0	54.3	-	72.9
	TM-RNN [40]	76.5	70.6	76.4	52.3	-	72.4
	DREAM [31]	84.8	76.1	81.6	55.1	-	78.3
	Joint-LSTM [34]	79.4	67.6	76.3	43.1	-	71.5
	M-BLSTM [19]	80.1	70.4	73.0	48.0	78.5	71.8
	PM-BLSTM [19]	81.6	71.3	74.4	48.6	78.9	73.0
	Att-BLSTM [41]	85.1	76.6	77.5	57.7	84.0	77.3
	BLSTML-SVM [42]	71.4	69.9	72.8	52.8	-	69.0
	Hierarchical BLSTMs [14]	81.9	77.4	78.0	58.4	-	78.5
	GRU [43]	-	-	-	-	-	72.2
	SGRU-CNN [17]	82.8	72.2	78.0	50.4	-	74.7
	UGC-DDI [44]	76.4	68.5	76.5	45.5	-	71.2
BERT	Basic-BERT [45]	-	-	-	-	-	79.9
	BioBERT [15]	86.1	80.1	84.6	56.6	-	80.9
	EMSI-BERT	86.8	80.7	86.6	56.0	88.0	82.0

*—No value is provided in the literature.

Firstly, the proposed method is compared with traditional statistical machine learning-based methods, including three baseline models (FBK-irst [27], WBI [28], and UTurku [29]), Kim et al.'s method [26], and the RBF-Linear method [30]. Comparison results are shown in Table 9. For all five methods, the SVM classifier is adopted for the interaction of the input drug entities, but they have significant differences in model structure, strategy, and feature selection. For the FBK-irst and Kim et al. methods, a binary classification model is firstly used to detect whether there is a relationship between a pair of entities, and then a multi-classification model is used to distinguish the specific class of an entity pair. In terms of the SVM model-based multi-classification strategy, the one-against-all strategy is adopted in FBK-irst method, while the one-against-one strategy is adopted in the Kim et al. method. Specially, WBI and UTurku, unlike Kim et al. and FBK-irst, directly used a multi-classification SVM to accomplish all tasks. In addition to the adjustment of the strategy, the kernel function design of the SVM can also affect the result of DDI, for example, the RBF-linear kernel function [30] achieves much better results than traditional SVM methods. However, as shown in Table 9, the proposed EMSI-BERT method significantly improves the F1-score of Advice, Effect, Mechanism, Int, and Dec compared with those traditional statistical machine learning-based DDI classification methods. For example, compared with the RBF-linear method currently offering a better result in traditional machine learning, the proposed EMSI-BERT method improved the micro-averaged F1-score from 71.1% to 82.0%. The classification effect also grows as the F1-score of each classification is improved by more than 10%. In addition, features still have an important role in the above five methods. Some commonly used features, such as n-gram, word pair, and part of speech information, are introduced into the models, and some higher-order features, such as syntax tree, dependence grammar, and syntax path information, are integrated into the model [27,29]. These traditional DDI classification methods manually integrating extracted features have obtained good results to some extent but have limited reusability and propagate errors due to a cascade of manually extracted features. As a result, the application of traditional DDI classification methods and the improvement of DDI extraction has been limited. The proposed deep learning-based DDI classification method, which directly uses an end-to-end approach for DDI extraction, reduces the complexity of manual feature extraction to a

certain extent, and avoids the error accumulation problem of cascaded multiple steps (such as syntactic analysis).

To evaluate the proposed EMSI-BERT, we sequentially compare it with CNN-, RNN- and BERT-related DDI classification models. In CNN-related models [10,35–38], related work first transforms the input text into word vectors, then uses the window convolution operation of CNN to achieve sequence modeling and feature extraction of entity pairs, and finally, an external classifier such as Softmax is utilized to classify the entity-relationship. The effects of classification of these methods are all improved based on the CNN network, such as changing a single-channel to a multi-channel CNN [36], combining typical convolutions and dilated convolutions to model far word distance [38], and integrating syntax information to modeling input sequence [10]. Unlike CNN-related models, RNN-related models such as Joint-LSTM [34], TM-RNN [40], DREAM [31], BLSTM [14,19,41,42], GRU [17,43], and UGC-DDI [44] directly perform sequence modeling. In order to enhance the effect of classification, there have been many improvements: (1) using an attention or memory mechanism to achieve soft sequence modeling; (2) using richer external information to achieve a better understanding, e.g., entity position information or general user context; (3) achieving a better understanding through the combination of LSTM, CNN, and SVM.

Considering that the limited domain-labeled data for DDI extraction cannot guarantee the sufficient model classification effect, we incorporate the asymmetrical Entity-Mask strategy and Symbol-Insert structure for DDI extraction. Compared with CNN- and RNN-related DDI classification methods, the micro-averaged F1-score of the proposed EMSI-BERT method has significant improvement, as shown in Table 10. Furthermore, Table 10 shows that the effect of the RNN-based sequence modeling methods is slightly better than that of the CNN-based methods, while the deep learning-based methods also have far better results than SVM-related models. It is worth noting that BERT-based methods [15,45] have also been used in DDI tasks recently. Table 10 shows the representative BERT-related work. Compared with Basic-BERT and BioBERT, the proposed EMSI-BERT has two different points: (1) the novel entity masking strategy for pre-training BERT is proposed. In contrast, other methods require a large amount of external information to achieve a better understanding. (2) The Symbol-Insert structure for fine-tuning BERT is designed to overcome the problem of different entity combinations sharing the same input sequence. In terms of accuracy, the proposed EMSI-BERT is also better than Basic-BERT and BioBERT.

The experimental results in Tables 9 and 10 illustrate that the EMSI-BERT method is superior to the SVM-, CNN-, RNN-, and BERT-related models. The proposed EMSI-BERT method obtains the highest micro-average F1-score and has a better classification effect for the four categories (Advice, Effect, Mechanism, and Int) than SVM-, CNN-, and RNN-related models. Moreover, the proposed EMSI-BERT method also obtains the best results for the two-classification task of drug relationship detection, and its micro-average F1-score exceeds that of Att-BLSTM by 4%.

5.3.3. Ablation Experiment

To further validate the effectiveness of the proposed pre-training BERT with the asymmetrical Entity-Mask strategy in this paper, we compare the effect of the Symbol-Insert structure with Basic-BERT initialization and Entity-Mask-BERT initialization. The experimental results in Table 11 show that the fine-tuning BERT under the asymmetrical Entity-Mask strategy initialization is better than that under Basic-BERT initialization, with its micro-average F1-score improved by 3.0%. In addition, the sensitivity of different methods to preprocessing rules is further analyzed in this paper. As shown in Table 12, the presence or absence of preprocessing has a 1.0% influence on the proposed EMSI-BERT. Before and after preprocessing, the difference with other models, especially CNN-related models, reaches about 4%. The findings indicate that the robustness of the proposed method is better and can effectively restrict the influence of "Other" samples.

Table 11. Comparison results with different model initialization.

Model Structure	Micro-Average F1-Score
Basic-BERT initialization+ Symbol-Insert structure	79.0
Entity-Mask-BERT+ Symbol-Insert structure (EMSI-BERT)	82.0

Table 12. Sensitivity of different methods to the preprocessing rule.

Method	No-Preprocessing	Preprocessing	Variation of Micro-Average F1-Score
CNN [35]	65.0	69.7	4.7
MCNN [36]	67.8	70.2	2.4
SCNN [37]	64.5	68.4	3.9
Joint-LSTM [34]	67.2	69.4	2.2
TM-RNN [40]	70.8	72.4	1.6
PM-BLSTM [19]	71.6	73.0	1.4
EMSI-BERT	81.0	82.0	1.0

5.3.4. Model Visualization

To further explore the depth features learned by EMSI-BERT and the influence weight of words, we visualize the attention weight between the current [CLS] node and other words on each layer of BERT, as shown in Figure 8. For convenient analysis and visualization, entity 1 and entity 2 are presented as A and B, respectively, and S1, E1, S2, and E2 are mapped to the special symbols [unseen1], [unseen2], [unseen3], and [unseen4] in the BERT word list, respectively. In Figure 8, the thickness of the lines represents the weight relation between the [CLS] node and other nodes. The following conclusions can be made qualitatively through visual weight analysis: (1) in shallow feature learning, EMSI-BERT mainly focuses on some features at the overall level of sentences. For example, the weights are mainly concentrated on the nodes [CLS] and [SEP] in layer0-layer7. (2) In the slightly higher-level feature learning, EMSI-BERT focuses on some detailed features at the non-overall level. For example, the weights are mainly concentrated on nodes other than [CLS] and [SEP] in layer8-layer11. (3) There are differences in each layer's focal features. For example, some entity boundary features are focused in layer6 and layer7, while some keywords containing entity relationships are focused, such as "increases". From these qualitative analyses, it can be seen that the proposed EMSI-BERT method can extract semantic information at different levels and granularities in continuous space.

The above result of performance evaluation, comparison and ablation experiments, and visualization illustrates that the proposed EMSI-BERT method has a comprehensive improvement in the two-classification task of DDI detection and the multi-classification task of DDI extraction. The improvement is mainly attributed to the following reasons:

(1) Compared with traditional machine learning-based methods, which measure semantics in discrete space and design handcrafted features, the proposed EMSI-BERT method introduces probability embedding to measure semantics in continuous space and uses the end-to-end approach for DDI extraction, thus reducing the complexity of manual feature extraction and the accumulation error of multiple steps.
(2) Compared with deep learning-based methods, such as BILSTM, CNN and BERT-related models, which are limited to the quality of the dataset and the amount of labeled data, the improved asymmetrical Entity-Mask strategy can compensate for the lack of entity orientation and retain entities' co-occurrence information on the basis of the idea of distance supervision. Ablation experiments show that the asymmetrical Entity-Mask strategy alleviates the problem of data sparsity and effectively improves the effect of downstream DDI classification.
(3) The Symbol-Insert structure, designed for fine-tuning BERT, overcomes the problem of different entity combinations sharing the same input sequence and achieves the end-to-end DDI extraction without destroying the structure of Entity-Mask-BERT.

The experimental results show that the designed structure can be adapted to the DDI extraction task effectively. Moreover, the visualization in Section 5.3.4 illustrates that Symbol-Insert-BERT can extract entity-level features, syntactic features, and semantic features for DDI extraction from shallow to deep layers.

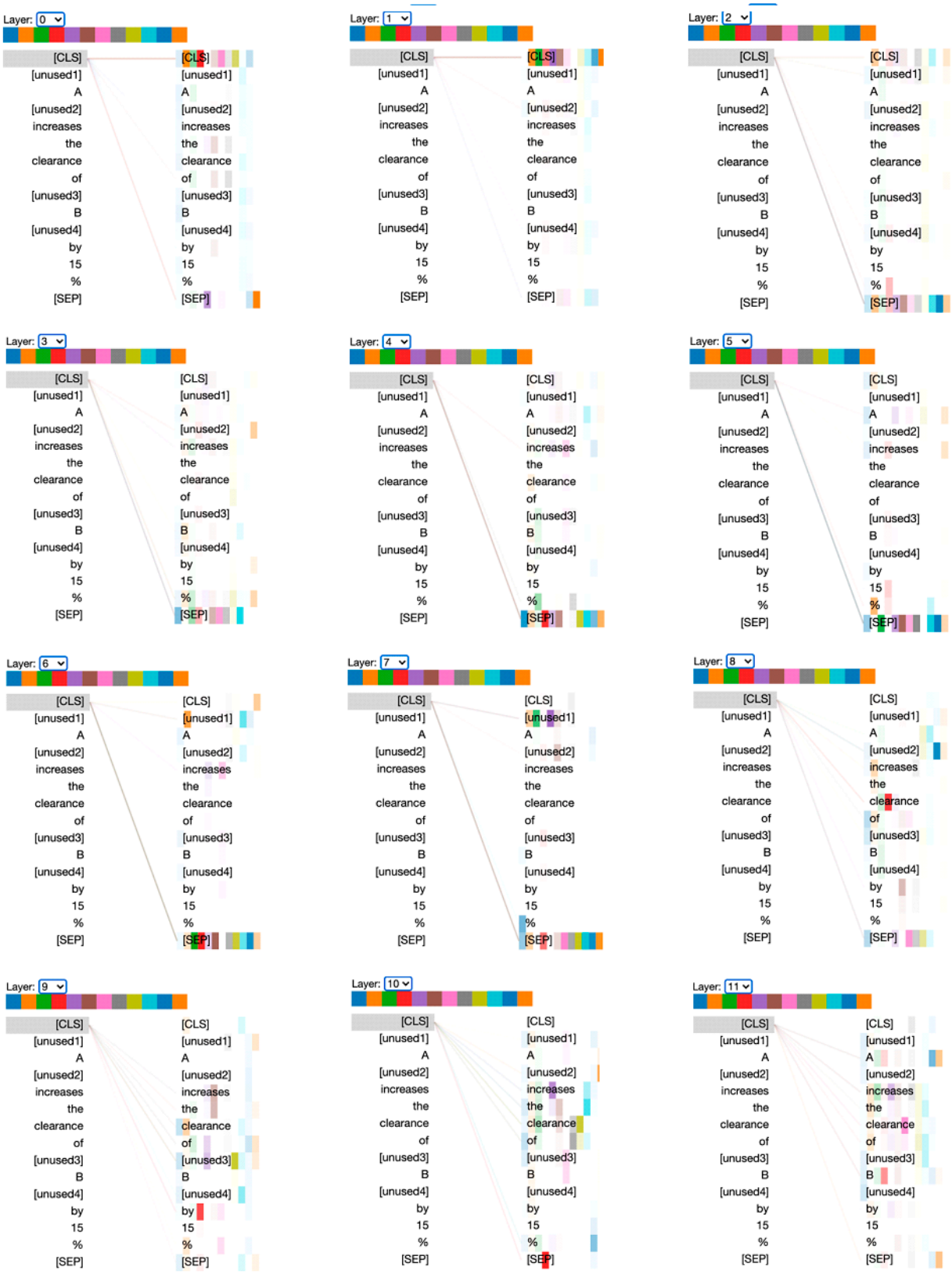

Figure 8. Attention weight visualization between the [CLS] node and other words.

6. Conclusions

Considering that discovering the drug–drug interaction relationship through medical experimentation requires a significant amount of human and material resources, EMSI-BERT is proposed for DDI extraction from the biomedical literature, which is a branch of artificial intelligence technology. To address the problems of limited domain-labeled data, weakly expressed co-occurring entities, and poor adaptation of downstream tasks in DDI extraction, the asymmetrical Entity-Mask strategy for pre-training BERT is improved and the Symbol-Insert structure for fine-tuning BERT is designed. In the pre- training stage of DDI extraction, the random masking approach in BERT is improved to the entity masking strategy, which preserved a certain amount of the co-occurring information of high frequency entities. The experimental results reveal that the proposed pre-training BERT with the asymmetrical Entity-Mask strategy can obtain better effects in downstream tasks than the baseline Basic-BERT. In the fine-tuning stage of DDI classification, the Symbol-Insert structure is designed to better adapt to the task of relational classification and retain as much parameter information of the pre-training BERT as possible. The experimental results on the DDI-Extraction 2013 dataset show that the proposed method comprehensively improved the multi-classification task of DDI extraction and the two-classification task of DDI detection. The proposed EMSI-BERT method needs to predict all combinations of entities, so its classification efficiency remains to be improved. Future studies should further address the pre-training strategy (such as the masking of other types of entities) and adjust the insertion way of positional symbols to account for both effect and efficiency. Moreover, the subsequent research will apply the proposed EMSI-BERT method to other biology-related fields, such as protein–protein relation extraction, and study its performance from different perspectives. Finally, we shall introduce fuzzy learning [56–58] to the proposed EMSI-BERT method, which may help make it more practical.

Author Contributions: Conceptualization, Z.H., N.A. and F.R.; methodology and investigation, Z.H., N.A. and J.L.; resources, Z.H. and F.R.; writing—original draft preparation, Z.H., N.A. and J.L.; writing—review and editing, Z.H., N.A., J.L. and F.R.; supervision, N.A. and F.R.; project administration, Z.H. and N.A.; funding acquisition, Z.H. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grant 62176084, the Natural Science Foundation of Anhui Province of China under grant nos. 1908085MF195 and 2008085MF193, the Key Project of Anhui University Excellent Young Talents Support Plan of China under grant gxyqZD2021122, the Natural Science Research Project of the Education Department of Anhui Province under grant 2022AH051038, and the Project of Provincial Key Laboratory for Computer Information Processing Technology of Soochow University of China under grant KJS1934.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under grant 62176084, the Natural Science Foundation of Anhui Province of China under grant nos. 1908085MF195 and 2008085MF193, the Key Project of Anhui University Excellent Young Talents Support Plan of China under grant gxyqZD2021122, the Natural Science Research Project of the Education Department of Anhui Province under grant 2022AH051038, and the Project of Provincial Key Laboratory for Computer Information Processing Technology of Soochow University of China under grant KJS1934. We acknowledge the use of the facilities and equipment provided by Hefei University of Technology. We would like to thank each contributor stated above for providing help and assistance in this research.

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

References

1. Qiu, Y.; Zhang, Y.; Deng, Y.; Liu, S.; Zhang, W. A Comprehensive review of Computational Methods for Drug-drug Interaction Detection. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2022**, *19*, 1968–1985. [CrossRef]
2. Thv, A.; Ngn, T.K.; Quk, H.; Le, N.Q.K. On the Road to Explainable AI in Drug-drug Interactions Prediction: A Systematic Review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2112–2123. [CrossRef]

3. Quan, C.; Wang, M.; Ren, F. An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature. *PLoS ONE* **2014**, *9*, e102039. [CrossRef]
4. Chen, J.; Bao, H.; Wei, P.; Qic, C.; Buz, T. Biomedical relation extraction via knowledge-enhanced reading comprehension. *BMC Bioinform.* **2022**, *23*, 20. [CrossRef] [PubMed]
5. Ibrahim, H.; Abdo, A.; El Kerdawy, A.M.; Eldin, A.S. Signal Detection in Pharmacovigilance: A Review of Informatics-driven Approaches for the Discovery of Drug-Drug Interaction Signals in Different Data Sources. *Artif. Intell. Life Sci.* **2021**, *1*, 100005. [CrossRef]
6. Feng, Y.-H.; Zhang, S.-W.; Zhang, Q.-Q.; Zhang, C.-H.; Shi, J.-Y. DeepMDDI: A Deep Graph Convolutional Network Framework for Multi-label Prediction of Drug-drug Interactions. *Anal. Biochem.* **2022**, *646*, 114631. [CrossRef] [PubMed]
7. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef]
8. Zanzoni, A.; Montecchi-Palazzi, L.; Quondam, M.; Ausiello, G.; Helmer-Citterich, M.; Cesareni, G. MINT: A Molecular INTeraction Database. *FEBS Lett.* **2002**, *513*, 135–140. [CrossRef]
9. Kerrien, S.; Aranda, B.; Breuza, L.; Bridge, A.; Broackes-Carter, F.; Chen, C.; Duesbury, M.; Dumousseau, M.; Feuermann, M.; Hinz, U.; et al. The IntAct Molecular Interaction Database in 2012. *Nucleic Acids Res.* **2011**, *40*, D841–D846. [CrossRef]
10. Park, C.; Park, J.; Park, S. AGCN: Attention-based Graph Convolutional Networks for Drug-drug Interaction Extraction. *Expert Syst. Appl.* **2020**, *159*, 113538–113550. [CrossRef]
11. Tran, T.; Kavuluru, R.; Kilicoglu, H. Attention-Gated Graph Convolutions for Extracting Drug Interaction Information from Drug Labels. *ACM Trans. Comput. Health* **2021**, *2*, 1–19. [CrossRef]
12. Zhu, J.; Liu, Y.; Wen, C.; Wu, X. DGDFS: Dependence Guided Discriminative Feature Selection for Predicting Adverse Drug-Drug Interaction. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 271–285. [CrossRef]
13. Zhu, J.; Liu, Y.; Zhang, Y.; Chen, Z.; She, K.; Tong, R.S. DAEM: Deep Attributed Embedding based Multi-task Learning for Predicting Adverse Drug–drug Interaction. *Expert Syst. Appl.* **2023**, *215*, 119312. [CrossRef]
14. Fatehifar, M.; Karshenas, H. Drug-Drug Interaction Extraction Using a Position and Similarity Fusion-based Attention Mechanism. *J. Biomed. Inform.* **2021**, *115*, 103707. [CrossRef]
15. Zhu, Y.; Li, L.; Lu, H.; Zhou, A.; Qin, X. Extracting Drug-drug Interactions from Texts with BioBERT and Multiple Entity-aware Attentions. *J. Biomed. Inform.* **2020**, *106*, 103451. [CrossRef]
16. Zaikis, D.; Vlahavas, I. TP-DDI: Transformer-based Pipeline for the Extraction of Drug-Drug Interactions. *Artif. Intell. Med.* **2021**, *119*, 102153. [CrossRef] [PubMed]
17. Wu, H.; Xing, Y.; Ge, W.; Liu, X. Drug-drug Interaction Extraction via Hybrid Neural Networks on Biomedical Literature. *J. Biomed. Inform.* **2020**, *106*, 103432. [CrossRef]
18. Zhang, Y.; Zheng, W.; Lin, H.; Wang, J.; Yang, Z.; Michel, D. Drug-drug Interaction Extraction via Hierarchical RNNs on Sequence and Shortest Dependency Paths. *Bioinformatics* **2018**, *34*, 828–835. [CrossRef] [PubMed]
19. Zhou, D.; Miao, L.; He, Y. Position-aware Deep Multi-task Learning for Drug–Drug Interaction Extraction. *Artif. Intell. Med.* **2018**, *87*, 1–8. [CrossRef] [PubMed]
20. Bunescu, R.; Mooney, R. Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline. In *Proceedings of the Hlt-Naacl Bionlp Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06*; Association for Computational Linguistics: New York, NY, USA, 2006; pp. 49–56. [CrossRef]
21. Fundel, K.; Küffner, R.; Zimmer, R. RelEx-Relation Extraction Using Dependency Parse Trees. *Bioinformatics* **2007**, *23*, 365–371. [CrossRef] [PubMed]
22. Segura-Bedmar, I.; Martínez, P.; de Pablo-Sánchez, C. A Linguistic Rule-based Approach to Extract Drug-drug Interactions from Pharmacological Documents. *BMC Bioinf.* **2011**, *12*, S1. [CrossRef]
23. An, N.; Xiao, Y.; Yuan, J.; Yang, J.; Alterovitz, G. Extracting Causal Relations from the Literature with Word Vector Mapping. *Comput. Biol. Med.* **2019**, *115*, 103524. [CrossRef]
24. Cui, B.; Lin, H.; Yang, Z. SVM-based Protein-Protein Interaction Extraction from Medline abstracts. In Proceedings of the 2007 Second International Conference on Bio-Inspired Computing: Theories and Applications, Zhengzhou, China, 14–17 September 2007; pp. 182–185. [CrossRef]
25. Segura-Bedmar, I.; Martínez, P.; de Pablo-Sánchez, C. Using a Shallow Linguistic Kernel for Drug–drug Interaction Extraction. *J. Biomed. Inform.* **2011**, *44*, 789–804. [CrossRef]
26. Kim, S.; Liu, H.; Yeganova, L.; Wilbur, W.J. Extracting Drug-drug Interactions from Literature Using a Rich Feature-based Linear Kernel Approach. *J. Biomed. Inform.* **2015**, *55*, 23–30. [CrossRef]
27. Chowdhury, M.F.M.; Lavelli, A. FBK-irst: A Multi-Phase Kernel Based Approach for Drug-Drug Interaction Detection and Classification that Exploits Linguistic Information. In Proceedings of the 7th International Workshop Semantic Evaluation, Atlanta, GA, USA, 14–15 June 2013; Volume 2, pp. 351–355.
28. Thomas, P.; Neves, M.; Rocktäschel, T.; Leser, U. WBI-DDI: Drug-Drug Interaction Extraction using Majority Voting. In Proceedings of the 2nd Joint Conference Lexical Computational Semantics, Atlanta, GA, USA, 13–14 June 2013; Volume 2, pp. 628–635.

29. Björne, J.; Kaewphan, S.; Salakoski, T. UTurku: Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. In Proceedings of the 2nd Joint Conference Lexical Computational Semantics, Atlanta, GA, USA, 13–14 June 2013; Volume 2, pp. 651–659.
30. Raihani, A.; Laachfoubi, N. Extracting Drug-drug Interactions from Biomedical Text Using a Feature-based Kernel Approach. *J. Theor. Appl. Inf. Technol.* **2016**, *92*, 109–120.
31. Shi, Y.; Quan, P.; Zhang, T.; Niu, L.F. DREAM: Drug-drug Interaction Extraction with Enhanced Dependency Graph and Attention Mechanism. *Methods* **2022**, *203*, 152–159. [CrossRef] [PubMed]
32. Chen, J.; Sun, X.; Jin, X.; Sutcliffe, R. Extracting Drug–drug Interactions from No-blinding Texts using Key Semantic Sentences and GHM Loss. *J. Biomed. Inform.* **2022**, *135*, 104192. [CrossRef] [PubMed]
33. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781v3.
34. Sahu, S.; Anand, A. Drug-drug Interaction Extraction from Biomedical Texts Using Long Short-term Memory Network. *J. Biomed. Inform.* **2018**, *86*, 15–24. [CrossRef]
35. Liu, S.; Tang, B.; Chen, Q.; Wang, X. Drug-Drug Interaction Extraction via Convolutional Neural Networks. *Comput. Math. Methods Med.* **2016**, *2016*, 6918381. [CrossRef] [PubMed]
36. Quan, C.; Hua, L.; Sun, X.; Bai, W. Multichannel Convolutional Neural Network for Biological Relation Extraction. *Biomed. Res. Int.* **2016**, *2016*, 1850404. [CrossRef] [PubMed]
37. Zhao, Z.; Yang, Z.; Luo, L.; Lin, H.; Wang, J. Drug Drug Interaction Extraction from Biomedical Literature using Syntax Convolutional Neural Network. *Bioinformatics* **2016**, *32*, 3444–3453. [CrossRef] [PubMed]
38. Sun, X.; Dong, K.; Ma, L.; Sutcliffe, R.; He, F.; Chen, S.; Feng, J. Drug-Drug Interaction Extraction via Recurrent Hybrid Convolutional Neural Networks with an Improved Focal Loss. *Entropy* **2019**, *21*, 37. [CrossRef] [PubMed]
39. Lim, S.; Lee, K.; Kang, J. Drug Drug Interaction Extraction from the Literature using a Recursive Neural Network. *PLoS ONE* **2018**, *13*, e0190926. [CrossRef] [PubMed]
40. Liu, J.; Huang, Z.; Ren, F.; Hua, L. Drug-Drug Interaction Extraction based on Transfer Weight Matrix and Memory Network. *IEEE Access* **2019**, *7*, 101260–101268. [CrossRef]
41. Zheng, W.; Lin, H.; Luo, L.; Zhao, Z.; Li, Z.; Zhang, Y.; Yang, Z.; Wang, J. An Attention-based Effective Neural Model for Drug-drug Interactions Extraction. *BMC Bioinf.* **2017**, *18*, 445. [CrossRef] [PubMed]
42. Huang, D.; Jiang, Z.; Li, Z.; Li, L. Drug–drug Interaction Extraction from Biomedical Literature Using Support Vector Machine and Long Short Term Memory Networks. *Inform. Sci.* **2017**, *415–416*, 100–109. [CrossRef]
43. Yi, Z.; Li, S.; Yu, J.; Wu, Q. Drug-drug Interaction Extraction via Recurrent Neural Network with Multiple Attention Layers. *Adv. Data Min. Appl.* **2017**, *10604*, 554–566. [CrossRef]
44. Xu, B.; Shi, X.; Yin, Y.; Zhao, Z.; Zheng, W.; Lin, H.; Yang, Z.; Wang, J.; Xia, F. Incorporating User Generated Content for Drug Drug Interaction Extraction Based on Full Attention Mechanism. *IEEE Trans. Nanobiosci.* **2019**, *18*, 360–367. [CrossRef]
45. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMO on Ten Benchmarking Datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 58–65. [CrossRef]
46. Peng, S.; Vijay-Shanker, K. Investigation of Improving the Pre-training and Fine-tuning of BERT model for Biomedical Relation Extraction. *BMC Bioinform.* **2022**, *23*, 120. [CrossRef]
47. Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237. Available online: https://aclanthology.org/N18-1202 (accessed on 10 May 2021).
48. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805v2.
49. Shah, S.M.A.; Y Ou, Y. TRP-BERT: Discrimination of Transient Receptor Potential (TRP) Channels using Contextual Representations from Deep Bidirectional Transformer based on BERT. *Comput. Biol. Med.* **2021**, *137*, 104821. [CrossRef]
50. Luzuriaga, J.; Munoz, E.; Rosales-Mendez, H.; Hogan, A. Merging Web Tables for Relation Extraction with Knowledge Graphs. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 1803–1816. [CrossRef]
51. Zhao, Q.; Xu, D.Z.; Li, J.Q.; Zhao, L.; Rajput, F.A. Knowledge Guided Distance Supervision for Biomedical Relation Extraction in Chinese Electronic Medical Records. *Expert Syst. Appl.* **2022**, *204*, 117606. [CrossRef]
52. GitHub. Available online: https://github.com/huangzhong3315/InsertBERT (accessed on 1 December 2022).
53. PubMed Dataset. Available online: https://www.nlm.nih.gov/databases/download/pubmed_medline.html (accessed on 23 March 2021).
54. PharmNet. Available online: http://www.pharmnet.com.cn/search/ (accessed on 2 May 2022).
55. Hugging Face. Available online: https://huggingface.co/bert-base-uncased (accessed on 1 May 2021).
56. Tang, Y.; Zhang, L.; Bao, G.; Ren, F.J.; Pedrycz, W. Symmetric Implicational Algorithm Derived from Intuitionistic Fuzzy Entropy. *Iran. J. Fuzzy Syst.* **2022**, *19*, 27–44. [CrossRef]

57. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Viewpoint-based Kernel Fuzzy Clustering with Weight Information Granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *in press*. [CrossRef]
58. Yang, J.Q.; Chen, C.H.; Li, J.Y.; Liu, D.; Li, T.; Zhan, Z.H. Compressed-Encoding Particle Swarm Optimization with Fuzzy Learning for Large-Scale Feature Selection. *Symmetry* **2022**, *14*, 1142. [CrossRef]