applied sciences

MDPI

Article

A Joint Domain-Specific Pre-Training Method Based on Data Enhancement

Yi Gan [1], Gaoyong Lu [1], Zhihui Su [1], Lei Wang [2,3], Junlin Zhou [3,4], Jiawei Jiang [3] and Duanbing Chen [3,4,*]

[1] The 10th Research Institute, China Electronic Technology Group Corporation, Chengdu 610036, China; sweat_gan@163.com (Y.G.); lugaoyong@163.com (G.L.); su-zhihui@163.com (Z.S.)
[2] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China; swjtuwanglei@gmail.com
[3] Chengdu Union Big Data Tech. Inc., Chengdu 610041, China; jlzhou@uestc.edu.cn (J.Z.); jiangjiawei@unionbigdata.com (J.J.)
[4] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China
* Correspondence: dbchen@uestc.edu.cn

Abstract: State-of-the-art performances for natural language processing tasks are achieved by supervised learning, specifically, by fine-tuning pre-trained language models such as BERT (Bidirectional Encoder Representation from Transformers). With increasingly accurate models, the size of the fine-tuned pre-training corpus is becoming larger and larger. However, very few studies have explored the selection of pre-training corpus. Therefore, this paper proposes a data enhancement-based domain pre-training method. At first, a pre-training task and a downstream fine-tuning task are jointly trained to alleviate the catastrophic forgetting problem generated by existing classical pre-training methods. Then, based on the hard-to-classify texts identified from downstream tasks' feedback, the pre-training corpus can be reconstructed by selecting the similar texts from it. The learning of the reconstructed pre-training corpus can deepen the model's understanding of undeterminable text expressions, thus enhancing the model's feature extraction ability for domain texts. Without any pre-processing of the pre-training corpus, the experiments are conducted for two tasks, named entity recognition (NER) and text classification (CLS). The results show that learning the domain corpus selected by the proposed method can supplement the model's understanding of domain-specific information and improve the performance of the basic pre-training model to achieve the best results compared with other benchmark methods.

Keywords: natural language processing; fine-tuning pre-training; data enhancement; joint training

check for updates

Citation: Gan, Y.; Lu, G.; Su, Z.; Wang, L.; Zhou, J.; Jiang, J.; Chen, D. A Joint Domain-Specific Pre-Training Method Based on Data Enhancement. *Appl. Sci.* **2023**, *13*, 4115. https://doi.org/10.3390/app13074115

Academic Editor: Habib Hamam

Received: 11 February 2023
Revised: 16 March 2023
Accepted: 21 March 2023
Published: 23 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background

In recent years, with the rapid development in the field of artificial intelligence, especially with the support of deep learning techniques, the development of Natural Language Processing (NLP) has made great progress under various tasks. Among these tasks, the development of pre-training techniques has played a crucial role. Pre-training models [1] provide an effective solution to the large-scale parameter learning problem in deep neural networks, which was first used in the field of Computer Vision (CV). The core idea is to pre-train deep neural networks on large datasets to obtain model parameters first, and then apply these trained models to various specific downstream tasks to avoid training from scratch and reduce the need for labeled data. Experiments show that pre-training on large corpora can significantly improve the performance of downstream tasks.

With the powerful generalization ability of pre-trained models, the idea of pre-training has been introduced into multiple fields to overcome research bottlenecks. In the field of speech recognition, Google [2] presented Wav2Vec 2.0, a framework for the self-supervised learning of speech representations that masks latent representations of the raw waveform

and solves a contrastive task over quantized speech representations. In highways research, as dynamic modulus testing is an expensive and complicated task that requires advanced testing equipment, researchers [3] used a deep convolutional neural network to predict a basic mechanical property (dynamic modulus) based on pre-training to reduce the laboratory effort needed. In addition, pre-trained models have been widely studied and applied in many fields such as computer vision [4], autonomous driving [5], and multimodal learning [6,7]. The emergence of pre-trained models has brought natural language processing into a new era.

In the field of natural language processing in recent years, Bidirectional Encoder Representation from Transformers (BERT) [8] and Generative Pre-trained Transformers (GPT) [9] are two language models that have attracted widespread attention. In particular, chatGPT, which is fine-tuned based on GPT3.5, has continued to gain popularity and become a hot topic. GPT [10] was developed based on the idea of generative pre-training and is mainly used to solve language generation tasks such as text generation and machine translation. It is suitable for pre-training on a large amount of unannotated data. In contrast, BERT is based on the idea of transfer learning and is more suitable for fine-tuning on annotated data. It is mainly used to solve language understanding tasks. The BERT model is regarded as a two-stage pre-training framework, including [11] (1) pre-train on a large unlabeled corpus and (2) fine-tune on task-specific labeled data. These models learn semantics during the pre-training step and require smaller labeled data to significantly improve their performance on downstream tasks. The fine-tuning step requires minor network modifications to create state-of-the-art models for various tasks.

Since the release of BERT, a two-stage approach based on "pre-training fine-tuning" has gradually become the mainstream of BERT-based natural language processing research. Currently, many scholars [12,13] have tried to improve the internal structure of the pre-training model substantially from transformer neural networks. In addition, by improving the two pre-training objectives of BERT [14,15], one can also greatly improve the learning ability of the model for text features. Furthermore, for domain-specific explicit knowledge, many scholars [16,17] have also started to investigate the method of incorporating external knowledge in the pre-training model, hoping to further enrich the text features learned by the model.

1.2. Motivation

The domain pre-training models with the mainstream approach of pre-training fine-tuning are generally performed on two different corpora, and aim to improve the robustness and generalization ability of the pre-trained models. Because the required corpus is very huge, the computational resources are highly consumed in the pre-training process. Therefore, in order to improve the effectiveness of the pre-trained models, the selection of the corpus for different tasks is an important research topic. Unfortunately, only a few researchers have investigated the problem of pre-training corpus selection. Inspired by above research, Ruder et al. [18] introduced a Bayesian optimization method to select from multiple data sources to obtain a high-quality pre-trained corpus. Learning by pre-training the screened corpus can effectively improve the robustness of different models on different domains and tasks. Furthermore, Salhofer et al. [19] proposed a sampling strategy which uses kernel density estimate to balance the instance selection between pseudo-perplexity and sentence length. The experiments show that the pre-trained models using the selected sample training set perform better, indicating that research targeting pre-training corpus selection can improve the accuracy of downstream tasks while reducing pre-training costs. Motivated by the aforementioned studies, this paper proposes a novel strategy for dynamically selecting pre-training corpora that not only enhances downstream task accuracy but also reduces the computational cost during the pre-training stage.

Besides the issue of high computational cost in pre-training, re-learning in downstream tasks can often cause the entire pre-trained model to forget the knowledge learned in the previous pre-training stage. To alleviate or avoid the catastrophic forgetting problem during

the fine-tuning process, Yang et al. [20] attempted to use the language model in the pre-training phase as an auxiliary objective function during fine-tuning, and to combine it with a task-specific optimization function. Phang et al. [21] designed Supplementary Training on Intermediate Labeled-data Tasks (STILTs) and found that additional supervised training was effective in improving the robustness of downstream tasks. This paper proposes a joint pre-training and fine-tuning model, while reducing the pre-training cost by adopting a dynamic corpus selection strategy. The proposed model alleviates catastrophic forgetting under the premise of reducing the pre-training cost.

1.3. Major Contribution

Above all, in order to reduce the pre-training cost and solve the catastrophic forgetting problem, a domain pre-training method based on data augmentation is proposed in this paper. This approach firstly combines the pre-training process with the downstream fine-tuning process to alleviate the catastrophic forgetting problem generated during the fine-tuning phase. Then, during the joint training process, the method reconstructs the pre-trained corpus by marking downstream task samples that the model judges to be wrong and selecting similar corpus samples from the domain corpus. The reconstructed pre-training corpus contains grammatical rules that cannot be captured by the pre-training model. Thus, by using the data augmentation-based joint training method, a smaller and higher quality pre-trained corpus can be automatically filtered without any pre-processing of the domain corpus, which can improve the adaptability of the model to a specific task and effectively alleviate the catastrophic forgetting problem. The contributions of the paper are as follows:

- A joint training strategy is proposed to capture the grammatical rules specific to the domain corpus and alleviate the catastrophic forgetting problem arising from the downstream fine-tuning process of traditional methods.
- Based on the joint training method, a data enhancement-based pre-training method is proposed to improve the feature extraction capability of the BERT model with hard-to-classify samples, making the pre-training process more automated, and avoiding the tedious data analysis and data processing work.
- The verification experiments on six datasets with three domains are conducted, which evaluate our data-enhancement-based pre-training method performance at two natural language process tasks (text classification and named entity recognition). The results show that our method can automatically complete the pre-training and fine-tuning task without data analysis and data pre-processing, and achieve the best results compared with the benchmark methods.

1.4. Organization

The remainder of the paper is organized as follows. The traditional pre-training mothod is introduced in Section 2 at first. The details of joint training strategy and data-enhancement-based pre-training method are addressed in Section 3. Experimental design and results analysis are discussed in Section 4. Finally, Section 5 concludes the work and presents the future directions.

2. Related Works

2.1. Traditional Pre-Training Procedure

The domain pre-training method based on the BERT model belongs to the unsupervised fine-tuning method, as shown in Figure 1a. The traditional pre-training strategy trains the domain pre-training model on a large amount of unlabeled data for various tasks firstly. This process consisted of two main semi-supervised tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM and NSP tasks can help the pre-trained model capture the representation of word-level and sentence-level, respectively. Since many studies [14,22–24] have questioned the effectiveness of the NSP task in BERT, this paper only takes MLM as the pre-training task. MLMs predict a "masked" word conditioned on

all other words in the sequence. Given a sequence of text $x_{(1:n)} = [x_1, x_2, \cdots, x_n]$, when training an MLM, words are chosen at random to be masked, using a special token [MASK], or replaced by a random token. The training objective is to recover the original tokens at the masked positions:

$$L_{MLM} = \sum_i m_i \cdot \log\left(P(x_i | x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n; \theta_T)\right), \quad (1)$$

where $m_i \in \{0, 1\}$ indicates whether x_i is masked or not, and θ_T are the parameters in the transformer encoder.

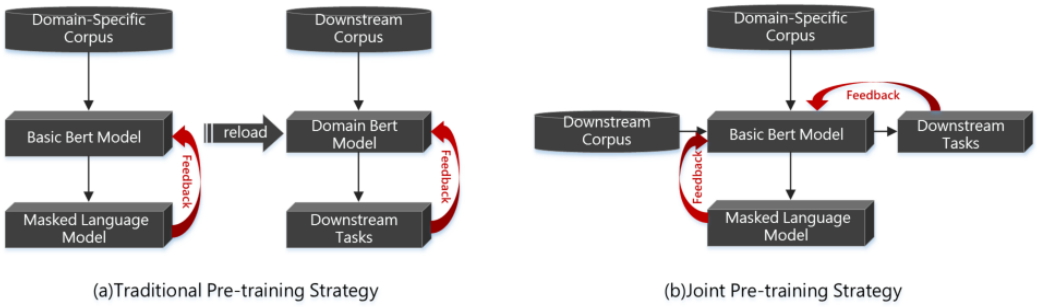


Figure 1. Traditional and Joint Pre-training Method.

A domain pre-trained model can be achieved through the pre-training process. Based on the specific downstream task, the fine-tuning process needs to add an output layer to the BERT model. After initializing the model with the pre-trained parameters, the parameters of the model are fine-tuned on the labeled data of the downstream task according to the target function L_{task} to exploit the pervasiveness of the pre-trained model better.

2.2. The Selection of Pre-Train Corpora

Currently, there are two strategies for selecting pre-train data: (1) collecting large, generic data, such as web crawls, news, and papers, with the goal of building universal language representations that can be useful across multiple domains; and (2) selecting in-domain corpora, which is often employed for domain-specific tasks. The first strategy is known to be computationally expensive and time-consuming, and the impact on downstream tasks remains uncertain. Consequently, for domain-specific tasks, researchers typically resort to the second strategy, which involves reconstructing a pre-train corpus using data specific to the domain. Recent studies on domain-specific BERT models, which are pre-trained on specialty source data, have empirically demonstrated that using in-domain data for pre-training can improve performance on downstream tasks [25–27].

However, in selecting in-domain corpora, intuition is often used, and the process remains an overlooked issue with only a limited amount of research dedicated to it. The first study on corpora selection [28] aimed to develop a cost-effective approach that could identify the most suitable source data for pre-training word vectors or language models from a range of options, based on a given named entity recognition (NER) dataset. The study used quantitative measures of similarity between pre-training corpora and downstream datasets, and found that these measures were predictive of the impact of pre-training data on final accuracy. Inspired by this study, Dai et al. [29] discovered that simple similarity measures could be used to nominate in-domain pre-training data. In addition to considering the overall similarity between pre-training corpora and downstream datasets, Salhofer et al. [19] proposed a sampling strategy that uses kernel density estimation to balance instance selection between pseudo-perplexity and sentence length.

3. Approach

The traditional pre-training procedure often suffers from two problems: (1) After fine-tuning on downstream tasks, pre-trained models often forget the knowledge learned during the pre-training phase, resulting in catastrophic forgetting. (2) Jointly training the pre-training process with downstream fine-tuning in Figure 1b is a solution that can effectively alleviate the catastrophic forgetting problem. However, due to the large size of pre-training corpus, it requires considerable computational resources and time, making joint training infeasible. Therefore, this paper proposes a data-enhancement-based domain pre-training method to address these two issues.

3.1. Joint Pre-Training Strategy

Existing traditional pre-training methods usually separate the pre-training phase from the downstream fine-tuning phase. As a result, this causes a catastrophic forgetting problem when the model is trained in the downstream fine-tuning task by forgetting the previous knowledge learned by the model in the domain corpus. To alleviate the catastrophic forgetting problem arising from the downstream fine-tuning process, this paper first proposes a joint training strategy that aims to capture the grammatical rules specific to the domain corpus using repeated learning of the domain corpus while learning a specific task, as shown in Figure 1b. The strategy first loads the domain pre-training corpus D_{domain} and the downstream task fine-tuning corpus D_{task} into the memory simultaneously. In each training episode, the update of the BERT parameters in the model is completed by calculating the MLM loss function L_{MLM} for each sample in D_{domain}. After learning all the pre-trained corpus, the downstream fine-tuning stage fine-tunes the parameters of the whole pre-trained model by calculating the loss L_{task} of each task sample for the downstream task corpus D_{task}. Thus, the language model in the pre-training phase is used as an objective function for information enhancement combined with the task-specific optimization function to obtain the joint training loss function, represented as follows:

$$L = \sum_{s \in D_{domain}} L_{MLM}(s) + \sum_{s \in D_{task}} L_{task}(s). \tag{2}$$

Such a joint training approach enables the model to perform the required downstream tasks well on the one hand, and on the other hand, allows the model to review the knowledge in the domain corpus while learning the knowledge of the downstream tasks. The joint learning of the corpus and the downstream task allows the BERT model to capture as many domain-unique expression patterns as possible to alleviate the catastrophic forgetting problem arising from traditional methods.

However, in such joint training, the size of the corpus data D_{domain} required for pre-training is very large and the quality of the data is uneven. Therefore, the joint training method can easily enable the model to learn the wrong information. At the same time, the pre-training in the joint training strategy leads to an increase in the computational resources required for model training because the pre-training corpus is so large, which makes it difficult to use the strategy for different downstream tasks and reduces the generalizability of the joint training method.

3.2. Joint Pre-Training Method Based on Data Enhancement

The domain expression is different from the real natural world expression. For example, the abstract of a paper tends to be more concise and highly general; the theoretical part tends to have stronger mathematical logic requirements, etc. In particular, different specific requirements of downstream tasks also affect the extraction of the required features by BERT. For example, there is a difference in the distribution of the domain corpus required for the classification tasks of abstract and conclusion and for the classification tasks of background and conclusion of a paper. Both the summary and conclusion of the paper have a high generality feature, so the feature of high generality cannot influence the judgment of this dichotomy. If the paper conclusion classification is replaced by the paper context

classification, high generality of the text is a major feature that can distinguish between these two types of texts. Therefore, the pre-training required precursors for the BERT model should depend on the specific data distribution of the downstream task.

In contrast, traditional pre-training methods often separate pre-training and downstream tasks, making the pre-training process heavily depend on the pre-trained corpus. This leads to the pre-training process learning wrong expressions or adding domain-irrelevant expressions into the model as long as there are dirty data or noise in the corpus. Therefore, existing pre-training processes often need to perform data analysis and data processing on the domain corpus prior to pre-training, and this work strongly relies on empirical knowledge.

In order to make pre-training more automated and avoid tedious data analysis, a data-enhancement-based pre-training method based on a joint training method is presented in this paper, as shown in Figure 2. In the joint training process, the corpus in the pre-training process is reconstructed for the corpus judged incorrectly in the downstream task. The text in the reconstructed pre-trained corpus has similar expressions (special word order, linguistic description logic, etc.) to the task text that the model cannot judge. Furthermore, pre-training tasks on such corpus will deepen the BERT model's understanding of these special expressions and enhance its feature extraction ability. For example, if BERT is unable to recognize the expressions described in the experimental conclusion of the paper, such as "*We observe that SCIBERT outperforms BERT-Base on biomedical tasks* (+1.92 *F*1 *with finetuning and* +3.59 *F*1 *without*)", It indicates that the current BERT model suffers from a lack of feature extraction capability for numerical-type experimental conclusion texts. At this time, we selected the similar corpus text with this numerical-type text from the corpus and reconstructed the pre-training corpus. The BERT pre-training continued under this corpus can strengthen the relevance of the BERT model with the feature extraction ability of this class of text and complement the comprehension ability of the BERT model for the expressions under this domain.

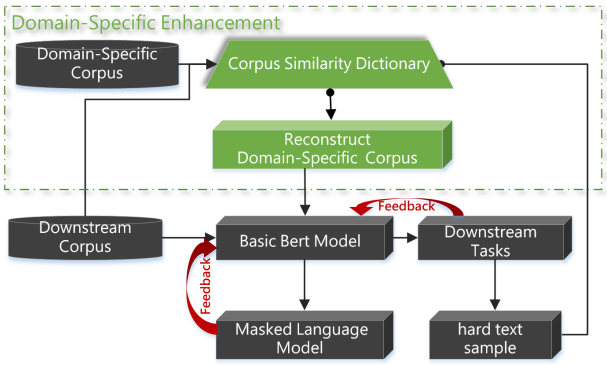


Figure 2. A Joint Pre-Training Method Based on Data Enhancement.

3.2.1. "Task2Domain" Similarity Dictionary

Generating a similarity dictionary between the domain corpus text and the downstream task text is an important part in the pre-training corpus reconstruction process stage, and the similarity of the text is the key of the process. Since the BERT model has powerful linguistic representation and feature extraction capabilities, the generic BERT model is used as a tool in this similarity calculation process. Firstly, for each sentence of the domain corpus and task text, the sentence is taken as input, and the first token position (CLS position) of the last output layer of the BERT model is selected as the representation of the sentence. For a given task text, the Euclidean distance between vectors is used as a measure of the similarity of two texts in order to find out which corpus texts are similar to that task text. It is worth noting that there are many methods [28,29] for calculating text similarity. This paper only uses the simplest text similarity calculation method as an example. Thus, the mapping between the task texts and similar corpus texts constitutes this similarity dictionary. In the similarity

dictionary establishment process, since the size of the domain corpus is often much larger than the size of the task text, it would take considerable computational resources to calculate the similarity between the each task text and each domain corpus. Therefore, in the process of selecting the $top-k$ corpus texts, in order to reduce the computation time, M corpus text is randomly selected as an alternative corpus in advance. Then by computing the similarity between that task text and the alternative corpus, the $=top-k$ corpus texts can be selected to establish this "Task2Domain" similarity dictionary $(T2D)$.

3.2.2. Model Optimization

Thus, the final loss function of the domain pre-training method based on data enhancement can be rewritten as follows:

$$L = \sum_{s \in D_{domain}{}'} L_{MLM}(s) + \sum_{s \in D_{task}} L_{task}(s), \tag{3}$$

where the domain corpus $D_{domain}{}'$ is dynamically changing in each training episode and the composition of $D_{domain}{}'$ strongly depends on the downstream task L_{task}.

To obtain the best pre-training model, the model typically undergoes multiple iterations of pre-training and fine-tuning training joint training. After the first round of downstream fine-tuning tasks, a set of hard-to-classify samples $D^t_{Err} = \{s_e | s_e \in D_{task}, f^t(s_e) == Error\}$ is obtained by identifying the samples that the current epoch(t)'s model $f^t(.)$ cannot classify correctly in downstream training set. Subsequently, using a pre-established "Task2Domain" similarity dictionary, we can obtain the $=top-k$ domain samples $Sim(s_e) = \{s | s \in T2D[S_e], s \in d_{domain}\}$ that have similar expressions with each hard-to-classify sample s_e. By aggregating the $top-k$ domain samples of all the hard-to-classify samples, we can obtain a smaller and higher quality pre-training corpus $D^t_{domain}{}' = \{s | s \in sim(s_e), s_e \in D_{Err}\}$ at epoch t. This approach makes the joint training of pre-training and fine-tuning feasible on the one hand, and reduces the influence of irrelevant samples and noise in pre-training on the model on the other hand.

The reconstruction of the in-domain corpus ($D^t_{domain}{}'$) for round t can provide an extra corpus for the next iteration of MLM pre-training stage. The main purpose of this round of pre-training is to supplement and strengthen the model's understanding of the knowledge that was lacking in the previous round of downstream tasks. Afterward, the model continues with the downstream fine-tuning tasks for round $t+1$. After multiple iterations, the model can retain as much knowledge as possible from the pre-training corpus while improving its accuracy on downstream tasks. The pseudocode for the entire training process using the joint pre-training method based on data enhancement can be found in Algorithm 1.

Algorithm 1: The Joint Pre-training Procedure Pseudocode.

Input: Domain Corpus (D_{domain}), Downstream Task Data (D_{task}),Training times (T) and Initial BERT.
Output:

```
Initialize BERT and downstream f(.) parameter.
Establishing "Task2Domain" similarity dictionary (T2D) according to BERT model based on D_domain and D_task.
for t from 0 to T do
    if t!=0 then
        Updating the BERT parameter according to L_MLM based on D^{t-1}_domain'.
    for s in D_task do
        Updating the downstream parameter f^t(.) according to L_task based on D_task' and marking the hard-to-classify sample s_e.
        Composing the hard-to-classify set as D^t_Err.
    Restructuring the pre-train corpus D^t_domain' based on the D^t_Err and T2D.
```

4. Experiments and Discussion

4.1. Corpora Datasets

This paper focuses on the construction and pre-training corpus of scientific and technical papers, and the pre-training corpus comes from two parts of all articles published in 2015 by ElsevierOA (https://elsevier.digitalcommonsdata.com/datasets/zm33cdndxs/2, accessed on 10 December 2022) and NIPS2015 (https://nips.cc/Conferences/2015, accessed on 10 December 2022) (Advances in Neural Information Processing Systems, a conference on advances in neural information processing systems). Among them, NIPS received a total of 1838 submissions and accepted 403 papers in 2015, with an acceptance rate of 21.9%. Among more than 400 papers, the directions are extremely diverse. The largest proportion of deep learning articles accounted for 11% of all accepted papers, followed by convex optimization accounting for 5% and statistical learning theory accounting for 3%. Furthermore, ElsevierOA [30] is a corpus of 40k (40,001) open access (OA) CC-BY articles from across Elsevier's journals representing the cross-discipline research of data at this scale to support NLP and ML research. This dataset was released to support the development of ML and NLP models targeting science articles from across all research domains. For these two corpus of papers, only the body text and abstract of each paper are selected as corpus sources to build this dataset.

4.2. Evaluation Tasks

In order to verify the effectiveness of the proposed method, two downstream tasks are selected: (1) named entity recognition (NER), and (2) text classification (CLS). NER is a span-level text classification task, which mainly refers to extracting entities with specific meanings or strong denotations from unstructured text. CLS is a sentence-level text classification task, which mainly refers to predicting text sentences into a specified category or categories.

To verify the effectiveness of the proposed method, text classification experiments on Sci-cite [31], Intent [32], and Chemprot [25] datasets were conducted first. In addition to the sentence-level classification task, this paper also performs the NER task on NCBI-disease [33], BC5CDR [34], and Sciie [35] datasets. The detail information about above datasets can be found in Table 1. For brevity, we do not explain the details of older datasets and readers can refer to the corresponding references.

Table 1. Statistics of the datasets.

Task	Dataset	Domain	#Label	#Entity	#Sents
CLS	Sci-cite	Multi	3	-	10,104
	Intent	CS	6	-	1941
	ChemProt	Bio	13	-	10,065
NER	NCBI-diseas	Bio	-	2	7287
	BC5CDR	Bio	-	2	15,030
	Sciie	CS	-	7	3187

Note: #Label and #Entity mean the number of types of Labels and Entities, #Sents means the number of sentences.

4.2.1. Text Classification Experiments

In addition to the basic BERT [8], two BERT-based models are compared in text classification: (1) Corpora-BERT pre-trained on above corpora, and (2) SciBERT [27] trained on the scientific literature text. Furthermore, the data enhancement joint pre-training method proposed in this paper based on the basic BERT model is called BERT-JDE. For each pre-trained model, we chose the base setting with 12 transformer layers and 768 hidden embedding sizes.

As shown in Table 2, the Corpora-BERT performs poorly. The reason is that Corpora-BERT learns from the raw corpora corpus during the pre-training process, which contains a large amount of noise and out-of-distribution domain corpus data. Without any pre-processing, the Corpora-BERT model finds it difficult to extract discriminating text features.

Therefore, even if fine-tuning is performed based on Corpora-BERT, the model still cannot achieve good downstream results. On the contrary, the BERT model trained by Google is based on a corpus with less noise and higher quality, which greatly enhances its feature extraction ability. After a small amount of fine-tuning, the pre-trained model can achieve better downstream results.

Table 2. Text Classification Experimental Results.

Dataset	BERT [8]		Corpora-BERT		SciBERT [27]		BERT-JDE	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Sci-cite	0.836	0.815	0.809	0.788	0.853	0.836	**0.857**	**0.850**
Intent	0.909	0.870	0.902	0.85	**0.920**	0.892	0.918	**0.893**
ChemProt	0.867	0.745	0.847	0.701	**0.902**	0.742	0.898	**0.801**

Furthermore, the BERT-JDE method proposed in this paper resamples a part of pre-trained samples through the feedback of the CLS task. It can enhance the classification ability of the model itself and improve the classification accuracy of the original BERT that was fine-tuned without any data analysis or processing of the pre-training sample. It indicates that by learning the domain pre-trained samples selected by this method, it can enhance the representation of the BERT model under the domain and provide more discriminating features for downstream tasks. Furthermore, compared with the SciBERT method, our BERT-JDE performs better in most cases. Because the corpus built in this paper covers a wide range of scientific and technological papers, BERT-JDE performs better on the Sci-cite and Intent datasets. However, although it performs more poorly than SciBERT on ChemProt dataset, it still achieves very competitive performance in BioText classification tasks.

4.2.2. Named Entity Recognition Experiment

In addition to above benchmarks, a data enhancement joint pre-training method with random sampling is specifically incorporated in named entity recognition, named BERT-RDE. This method randomly selects equal amounts of corpus data to constitute a pre-training corpus during each training round.

The experimental results of macro-F1 scores are shown in Table 3. The Corpora-BERT method is less effective than the original BERT model in BioText named entity recognition tasks. The reason is that the corpora constructed in this paper lacks special biological expressions and proper nouns. Therefore, some expressions of CS are noise for this task, which leads to poor performance of Corpora-BERT.

Table 3. Named Entity Recognition Experimental Results.

Dataset	Sciie	NCBI-Disease	BC5DR	Average
BERT [8]	0.652	0.886	0.862	0.800
SciBERT [27]	0.729	**0.925**	**0.919**	0.858
Corpora-BERT	0.671	0.879	0.863	0.804
BERT-RDE	0.694	0.893	0.88	0.822
BERT-JDE	0.713	0.895	0.918	0.842
SciBERT-JDE	**0.749**	**0.925**	**0.919**	**0.864**

Compared with the basic BERT model, the BERT-RDE method is a joint pre-training method which randomly selects 2000 texts at each round as the pre-training corpus. Although the BERT-RDE method reduces the corpus size, it has produced positive benefits (improvements of around 2.2%) compared with BERT model. This is because the randomly selected corpora of BERT-RDE can reduce the impact noise in pre-training corpus and enhance the generalization ability of the joint pre-training model, thereby avoiding over-fitting.

Compared with the basic BERT model, the BERT-JDE model proposed in this paper achieved a significant improvement of 5% on average, which shows the effectiveness of our data enhancement strategy. However, compared with the SciBERT model, the BERT-JDE method is slightly inferior. The reason is that the SciBERT model has a stronger feature extraction ability in the biological text NER task than the BERT model. Therefore, this experiment conducted a joint training of data enhancement on SciBERT and obtained SciBERT-JDE. The experiment shows that SciBERT-JDE achieves the best effects on the task of NER in multi-disciplinary papers (Sciie). On the BioText datasets (NCBI-disease and BC5DR), the SciBERT-JDE method is not worse even though the corpus lacks biological knowledge. Overall, the data enhancement strategy proposed in this paper still works for SciBERT, and the average entity recognition accuracy of SciBERT-JDE has improved.

Based on above experiments, it can be observed that the joint pre-training method based on data augmentation is effective for most BERT-based pre-training models. Without any pre-processing or analysis of the pre-training corpus, the proposed joint pre-training strategy and data augmentation method can further improve the performance of the pre-trained BERT-based pre-training model on downstream tasks. This approach achieves data augmentation through joint training with feedback from downstream tasks, which not only alleviates the catastrophic forgetting problem caused by downstream fine-tuning tasks, but also achieves the best results in many indicators. It has strong practical value as it achieves promising results without significantly increasing the time complexity and hardware requirements of the model training.

4.3. Ablation Experiments

In order to verify whether the data enhancement strategy proposed in this paper is effective, we removed the data enhancement module and kept only the joint training method, named BERT-J. Because the size of the domain corpus in the pre-training phase of the joint training process is very large (more than 1.19 GB), considerable computational resources are required if multiple pre-training rounds are performed. Therefore, only two datasets (Sci-cite and Intent) were selected for the ablation experiments, and the text classification experimental results are shown in Table 4.

Table 4. Comparison among data-enhancement strategies.

Dataset	BERT-J		BERT-JDE	
	ACC	**F1**	**ACC**	**F1**
Sci-cite	0.812	0.787	0.847	0.828
Intent	0.905	0.851	0.918	0.893

Compared with the BERT-J method without data augmentation, the BERT-JDE method enhances the understanding of the BERT model for domain-specific, model-unrecognizable expressions by reconstructing the pre-trained model using unrecognizable task text. In the text classification task, the domain corpus-enhanced BERT model has a more forward feature extraction capability to extract more discriminating text features under the domain restriction, and a stronger understanding of domain expression features. This leads to more significant extraction for the downstream text classification task. It indicates that the corpus reconstruction strategy with downstream feedback can greatly improve the model's effectiveness for the downstream task.

4.4. Parameter and Time Complex Analysis

In this section, we present a detailed analysis of the impact of the hyperparameter $top - k$ on the performance and time complexity of our proposed model. The range of $top - k$ was set from 0 to 500, and experiments were conducted on two popular natural language processing tasks: text classification(CLS) using Intent dataset and named entity recognition (NER) using the NCBI-disease and Sciie datasets. Furthermore, other hyperparameters set

as $M = 2000$ and $epoch = 20$. Particularly, when $top - k$ is set to 0, the model degenerates to the traditional method. The primary goal of this analysis is to investigate how $top - k$ affects the model's accuracy and efficiency, providing insights into the optimal setting of this hyperparameter. Furthermore, this study aims to provide a comprehensive evaluation of the proposed model's practical applicability for real-world NLP tasks. The results of the experiments are shown in Figure 3.

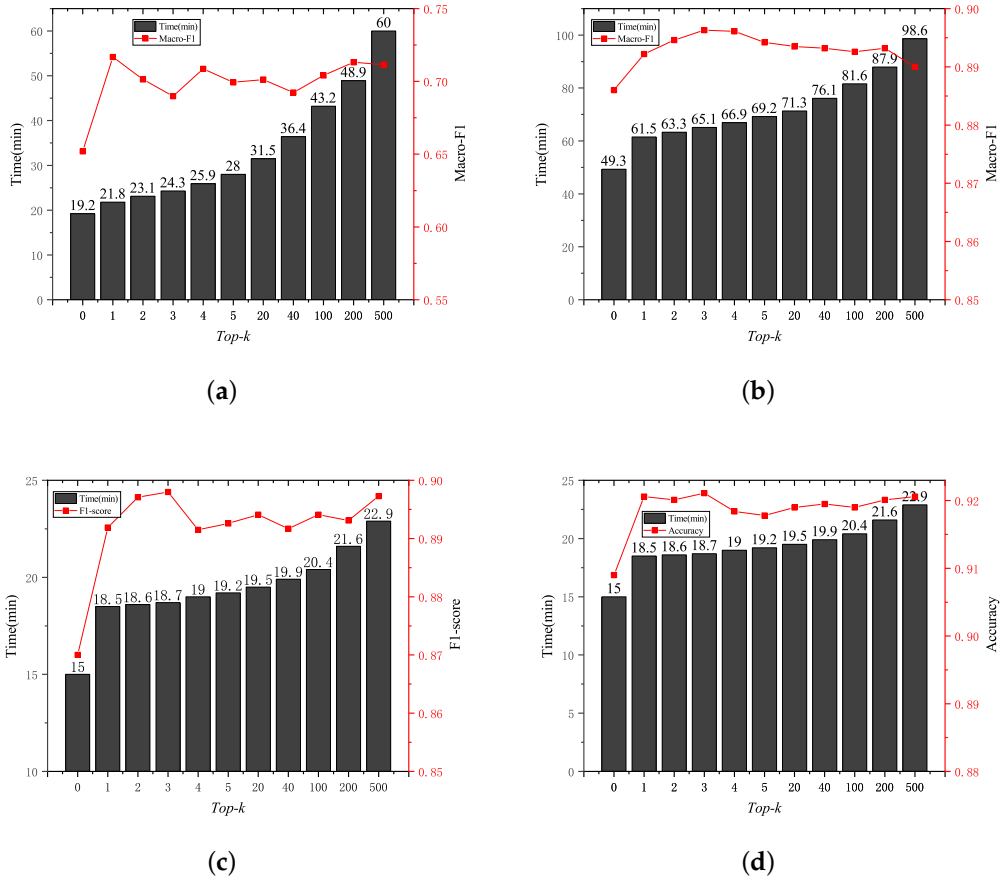


Figure 3. Parameter and time complex analysis of varying the $top - k$ for BERT-JDE. (**a**) NER performance on Sciie. (**b**) NER performance on NCBI-disease. (**c**) CLS performance of F1-score on Intent. (**d**) CLS performance of Accuracy on Intent.

4.4.1. Parameter Analysis

The results shown in Figure 3 indicate that the $top - k$ has little effect on the model performance. This is because the selected $top - k$ domain-specific text corpora for a hard-to-classify task sample have a consistent theme and characteristics, which limits the knowledge that can be learned from the reconstructed domain-specific text corpora. Therefore, increasing the $top - k$ parameter beyond a certain point does not proportionally increase the amount of information in the reconstructed domain-specific text corpora, while a larger $top - k$ may increase the size of the reconstructed domain-specific text corpora, it does not necessarily lead to a significant increase in information. On the other hand, setting the parameter $top - k$ too large can significantly increase the probability of introducing noisy and irrelevant corpora, which may lead to a decline in model performance. If $top - k$ is 0, "Task2Domain" similarity dictionary is empty; this indicates that data augmentation is not performed. This is equivalent to the traditional two-stage pre-training method, and the performance is relatively poor (see BERT in Tables 2 and 3).

In summary, we recommend selecting a moderate value of $top - k$ to achieve optimal performance. This value should be carefully chosen to balance the trade-off between incorporating new knowledge and avoiding the introduction of noise and irrelevant data.

4.4.2. Time Complex Analysis

This subsection presents an analysis of the time complexity of BERT-JDE with the $top-k$ parameter, which specifies the number of top items to be returned. During the data augmentation process, the size of the reconstructed pre-training corpus is directly related to the $top-k$. A larger $top-k$ means a larger 'Task2Domain' dictionary, which leads to higher time complexity for each round of pre-training corpus reconstruction and model training.

From Figure 3, it can be observed that as $top-k$ increases, the time complexity of the pre-training process exponentially grows, while the model performance only slightly fluctuates. This result confirms that $top-k$ significantly affects the training speed of the model. Particularly, when $top-k=0$, the model degenerates to the traditional method. Although the training time is reduced, the pre-training model without data augmentation strategy experiences a significant drop in downstream task performance (decreased by 2.2% on the Sciie NER experiment). This indicates that it is essential to incorporate a data augmentation strategy based on the $top-k$ in pre-training, which can significantly enhance the performance of the pre-training model on downstream tasks.

5. Conclusions

In order to make up for the lack of research on corpus sample selection in the current pre-training field, this paper proposes a novel joint pre-training method based on data enhancement.The method reconstructs the pre-trained corpus by selecting the approximate text from the pre-trained corpus for the downstream task of judging the incorrect task text. The method can correct the pre-training BERT's understanding of features such as judgment error text expressions and linguistic description logic, so as to improve the accuracy of downstream tasks. The experiments show that without any data analysis or data processing, our method can achieve the improvement of the basic BERT model based on any raw domain corpus. From the practical perspectives, since the corpus construction in this study mainly relied on academic research papers in the field of computer science, this limited the performance of the model in the biology domain. To improve the performance of the joint training model proposed in this study, a more extensive and higher quality pre-training corpus needs to be developed in the future. Therefore, how to build and expand a more universal pre-training corpus is one of the important directions that we need to study in the future. From the academic perspectives, this study only used the BERT representations as features to construct a dictionary and selected similar samples from the pre-training corpus. If the similarity between texts can be measured from multiple aspects, such as grammar, sentiment, or keywords, it would facilitate the development of a higher quality pre-training corpus and enable the attainment of a pre-training model with better performance.

Author Contributions: Conceptualization, Y.G. and L.W.; Methodology, Y.G., L.W. and D.C.; Software, G.L.; Validation, J.Z.; Formal analysis, Z.S.; Investigation, Y.G. and D.C.; Resources, G.L. and Z.S.; Data curation, G.L. and J.J.; Writing—original draft, L.W.; Writing—review & editing, G.L., Z.S., J.Z., J.J. and D.C.; Visualization, Z.S.; Supervision, D.C.; Project administration, J.J.; Funding acquisition, D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Program of National Natural Science Foundation of China with Grant No T2293771, by Key Research Project on Philosophy and Social Sciences of the Ministry of Education under Grant No 21JZD055 and by the Fundamental Research for the Central Universities with Grant No ZYGX2019J074.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Our data is actually open source data, the specific address is as follows: For ElsevierOA, we can get from https://elsevier.digitalcommonsdata.com/datasets/zm33cdndxs/2. For NIPS2015, we can get from https://nips.cc/Conferences/2015.

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

References

1. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Zhang, L.; Han, W.; Huang, M.; et al. Pre-Trained Models: Past, Present and Future. *AI Open.* **2021**, *2*, 225–250. [CrossRef]
2. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
3. Moussa, G.S.; Owais, M. Pre-trained deep learning for hot-mix asphalt dynamic modulus prediction with laboratory effort reduction. *Constr. Build. Mater.* **2020**, *265*, 120239. [CrossRef]
4. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
5. Wu, P.; Chen, L.; Li, H.; Jia, X.; Yan, J.; Qiao, Y. Policy Pre-training for End-to-end Autonomous Driving via Self-supervised Geometric Modeling. *arXiv* **2023**, arXiv:2301.01006.
6. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; pp. 2592–2607.
7. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
9. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 1877–1901.
10. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
11. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual Event, 5–10 July 2020; pp. 8342–8360.
12. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 5926–5936.
13. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 5753–5763.
14. Joshi, M.; Chen, D.; Liu, Y.; Weld, D.S.; Zettlemoyer, L.; Levy, O. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 64–77. [CrossRef]
15. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
16. Peters, M.E.; Neumann, M.; Logan, R.; Schwartz, R.; Joshi, V.; Singh, S.; Smith, N.A. Knowledge Enhanced Contextual Word Representations. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019.
17. Wang, R.; Tang, D.; Duan, N.; Wei, Z.; Huang, X.J.; Ji, J.; Cao, G.; Jiang, D.; Zhou, M. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; pp. 1405–1418.
18. Ruder, S.; Plank, B. Learning to select data for transfer learning with Bayesian Optimization. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 372–382.
19. Salhofer, E.; Liu, X.L.; Kern, R. Impact of Training Instance Selection on Domain-Specific Entity Extraction using BERT. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, Online, 10–15 July 2022; pp. 83–88.
20. Yang, J.; Zhao, H. Deepening Hidden Representations from Pre-trained Language Models. *arXiv* **2019**, arXiv:1911.01940.
21. Phang, J.; Févry, T.; Bowman, S.R. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv* **2018**, arXiv:1811.01088.
22. Hao, Y.; Dong, L.; Wei, F.; Xu, K. Visualizing and Understanding the Effectiveness of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4143–4152.
23. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
24. 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.

25. 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. [CrossRef] [PubMed]
26. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78.
27. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
28. Dai, X.; Karimi, S.; Hachey, B.; Paris, C. Using Similarity Measures to Select Pretraining Data for NER. *arXiv* **2019**, arXiv:1904.00585.
29. Dai, X.; Karimi, S.; Hachey, B.; Paris, C. Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Virtual Conference, 16–20 November 2020; pp. 1675–1681.
30. Kershaw, D.; Koeling, R. Elsevier oa cc-by corpus. *arXiv* **2020**, arXiv:2008.00774.
31. Cohan, A.; Ammar, W.; van Zuylen, M.; Cady, F. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 3586–3596.
32. Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; Jurafsky, D. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 391–406. [CrossRef]
33. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef] [PubMed]
34. 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] [PubMed]
35. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3219–3232.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.